# Analyse du Comportement Humain à Partir de la Vidéo en Étudiant l'Orientation du Mouvement

## THÈSE

présentée et soutenue publiquement le 19 novembre 2012

pour l'obtention du

**Doctorat de l'Université des Sciences et Technologies de Lille**

**(spécialité informatique)**

par

Yassine Benabbas

**Composition du jury**

| | | |
|---|---|---|
| *Président :* | Rémi Gilleron, (Professeur) | Université Lille 3 |
| *Rapporteurs :* | Bernard Merialdo, (Professeur) | Eurecom Sophia Antipolis |
| | Cyril Carincotte, (HDR) | Institut MULTITEL Mons |
| *Examinateur :* | Peter Veelaert, (Professeur) | Université de Gand |
| *Directeur de thèse :* | Chabane Djeraba, (Professeur) | Université Lille 1 |

**UNIVERSITÉ DES SCIENCES ET TECHNOLOGIES DE LILLE**
Laboratoire d'Informatique Fondamentale de Lille — UPRESA 8022
U.F.R. d'I.E.E.A. – Bât. M3 – 59655 VILLENEUVE D'ASCQ CEDEX
Tél. : +33 (0)3 28 77 85 41 – Télécopie : +33 (0)3 28 77 85 37 – email : direction@lifl.fr

Mis en page et composé avec LaTeX et la classe thlifl.

*À mes parents...*

*À mes frères...*

*À toute ma famille...*

*À tous mes amis...*

# Remerciements

Tout d'abord, je remercie vivement le professeur Chabane Djeraba de m'avoir accueilli au sein de son équipe pour réaliser ma thèse. Je le remercie pour m'avoir guidé, conseillé et soutenu tout au long de ce travail.

Je suis très reconnaissant envers les professeurs Dan Simovici, Slimane Larabi et Mark Zhang pour leurs discussions, leurs idées et leurs disponibilités. Mes remerciements vont également à Adel Lablack, à Nacim Ihaddadene, à Thierry Urruty et à Tarek Yahiaoui pour leur collaboration, leur aide et leur soutien pendant ces trois années.

Je tiens à remercier également les professeurs Bernard Merialdo et Cyril Carincotte pour avoir accepté d'être les rapporteurs de mon mémoire de thèse, et le professeur Peter Veelaert pour avoir accepté d'être examinateur.

Je remercie vivement Rémi Gilleron, professeur à l'Université Lille3, d'être le Président de mon jury de thèse. Je remercie aussi l'ensemble des membres du LIFL, doctorants, chercheurs et personnels administratifs qui ont contribué de près ou de loin à ma thèse.

Je tiens à saluer l'ensemble de l'équipe Fox-Miire dans laquelle j'ai réalisé ma thèse. Plus particulièrement, Jean et Marius pour l'aide apportée durant la rédaction de ce mémoire, ainsi que les autres doctorants et post-doctorants Samir, Ismail, Haidar, Taner, Rémy et Amel.

# Abstract

The recognition and prediction of people activities from videos are major concerns in the field of computer vision.

The main objective of my thesis is to propose algorithms that analyze human behavior from video. This problem is also called video content analysis or VCA. This analysis is performed in outdoor or indoor environments using simple webcams or more sophisticated surveillance cameras.

The video scene can be of two types depending on the number of people present. The first type is characterized by the presence of only one person at a time in the video. We call this an individual scene where we will tackle the problem of human action recognition. The second type of scene contains a large number of persons. This is called a crowd scene where we will address the problems of motion pattern extraction, crowd event detection and people counting.

To achieve our goals, we propose an approach based on three levels of analysis. The first level is the detection of low-level descriptors retrieved from the images of the video (e.g. areas in motion). The second level retrieves descriptors for modeling human behavior (e.g. average speed and direction of movement). The top level uses the descriptors of the intermediate step to provide users with concrete results on the analysis of behavior (e.g. this person is running, that one is walking, etc.).

Experimentation on well-known benchmarks have validated our approaches, with very satisfying results compared to the state of the art.



**Keywords : Computer vision, classification, event detection, action recognition, counting, motion patterns.**

# Résumé

La reconnaissance du comportement et la prédiction des activités des personnes depuis la vidéo sont des préoccupations majeures dans le domaine de la vision par ordinateur.

L'objectif principal de mon travail de thèse est de proposer des algorithmes qui permettent d'analyser des objets en mouvement à partir de la vidéo pour extraire des comportements humains. Cette analyse est effectuée dans des environnements intérieurs ou extérieurs filmés par de simples webcams ou par des caméras plus sophistiquées.

La scène analysée peut être de deux types en fonction du nombre de personnes présentes. On distingue les scènes de foule où le nombre de personnes est important. Dans ce type de scène, nous nous intéressons aux problèmes de la détection d'évènements de foule, à l'analyse des flux et à l'extraction des motifs de mouvement. Le deuxième type de scène se caractérise par la présence d'une seule personne à la fois dans le champ de la caméra. Ce type de scène est appelé scène individuelle. Nous y traitons le problème de reconnaissance d'actions humaines.

Pour atteindre ces objectifs, nous proposons une approche basée sur trois niveaux d'analyse. Le premier est l'extraction des caractéristiques de bas niveau depuis les images constituant un flux vidéo (ex. les zones en mouvement). Le deuxième construit des descripteurs pour l'analyse du comportement humain (ex. la direction moyenne et la vitesse moyenne de mouvement). Le niveau le plus haut se sert des descripteurs du deuxième niveau afin de fournir aux utilisateurs des résultats concrets sur l'analyse du comportement humain (ex. telle personne marche, une autre court, etc.).

Des expérimentations sur des benchmarks connus ont validé nos approches, avec un positionnement qui nous semble intéressant par rapport à l'état de l'art.

**Mots clés : Analyse du comportement, vision par ordinateur, classification, détection d'évènements, reconnaissance d'actions, comptage, motifs de mouvement.**

# Table des matières

**Table des figures** **xi**

**Liste des tableaux** **xvii**

**1 Introduction** **1**

1.1 Contexte . . . . . 2
1.2 Problématique . . . . . 8
1.3 Objectifs . . . . . 9
1.4 Contribution et originalité . . . . . 10
1.5 Schéma général de l'approche . . . . . 11
1.6 Plan de la thèse . . . . . 13

**2 État de l'art** **15**

2.1 Introduction . . . . . 16
2.2 Définition des zones d'intérêt . . . . . 16
2.2.1 Zones prédéfinies . . . . . 17
2.2.2 Avant-plan extrait par une méthode d'extraction d'arrière-plan . . . . . 18
2.2.3 Points d'intérêt . . . . . 20
2.2.4 Synthèse . . . . . 23
2.3 Descripteurs de mouvement . . . . . 25
2.3.1 Flux optique . . . . . 25
2.3.2 Volumes spatiotemporels . . . . . 30
2.3.3 Les images d'historique et d'énergie du mouvement . . . . . 31
2.3.4 Synthèse . . . . . 33
2.4 Méthodes de classification basée sur l'apprentissage automatique . . . . . 34
2.4.1 Les Machines à Vecteurs de Support (SVM) . . . . . 34
2.4.2 AdaBoost . . . . . 35

2.5 Analyse du comportement humain dans des scènes individuelles . . . . . 35
2.5.1 Représentations globales . . . . . 37
2.5.2 Représentations locales . . . . . 38
2.5.3 Synthèse . . . . . 41
2.6 Analyse du comportement humain dans des scènes de foule . . . . . 42
2.6.1 Extraction des motifs de mouvement . . . . . 42
2.6.2 Détection d'évènements de foule . . . . . 49
2.6.3 Estimation des flux . . . . . 54
2.7 Conclusion . . . . . 66

**3 Descripteurs pour l'analyse du comportement humain 69**
3.1 Introduction . . . . . 70
3.2 Modèle directionnel et modèle de magnitude . . . . . 70
3.3 Descripteurs sur les groupes de personnes en mouvement . . . . . 76
3.3.1 Extraction des blobs depuis une carte spatiotemporelle 2D . . . . . 76
3.3.2 Groupes de personnes dans une représentation sous forme de grille . . . 79
3.4 Conclusion . . . . . 81

**4 Analyse du comportement humain dans des scènes individuelles 83**
4.1 Introduction . . . . . 84
4.2 Description de l'approche . . . . . 84
4.3 Reconnaissance de l'action (niveau sémantique) . . . . . 86
4.4 Expérimentations et résultats . . . . . 88
4.4.1 Efficacité de la reconnaissance des actions . . . . . 88
4.4.2 Étude comparative . . . . . 90
4.4.3 Étude du nombre de classes d'actions et de la taille des blocs . . . . . 92
4.5 Conclusion . . . . . 92

**5 Analyse du comportement humain dans une scène de foule 95**
5.1 Extraction des motifs de mouvement . . . . . 96
5.1.1 Description de l'approche . . . . . 97
5.1.2 Extraction des motifs de mouvement (niveau sémantique) . . . . . 98
5.1.3 Expérimentations . . . . . 100
5.2 Détection d'évènements de foule . . . . . 105
5.2.1 Description de l'approche . . . . . 105

5.2.2 Détection d'évènements de foule (niveau sémantique) . . . . . 106
5.2.3 Expériementations et résultats . . . . . 111
5.3 Estimation des flux . . . . . 116
5.3.1 Description de l'approche . . . . . 117
5.3.2 Décision de comptage (niveau sémantique) . . . . . 118
5.3.3 Expériences et résultats . . . . . 121
5.4 Conclusion . . . . . 125

**6 Conclusions et perspectives** **127**
6.1 Introduction . . . . . 128
6.2 Résumé de nos contributions . . . . . 128
6.3 Travaux futurs . . . . . 130

**7 Publications** **133**
7.1 Livres et revues . . . . . 133
7.2 Chapitres de livres . . . . . 133
7.3 Conférences Internationales . . . . . 134
7.4 Conférences Nationales . . . . . 134

**Bibliographie** **137**

**A Définitions** **155**

**B Description de MIAUCE** **159**

**C Description du projet CAnADA** **163**

# Table des figures


1.1 Illustration de quelques applications de l'analyse du comportement humain depuis la vidéo . . . . . 4

1.2 Illustration d'un système de vidéo-surveillance automatique . . . . . 5

1.3 Schéma d'un système de vidéo-surveillance de troisième génération typique . . 7

1.4 Illustration de notre approche globale pour l'analyse du comportement humain 12

2.1 Accumulation de la ligne virtuelle à travers le temps . . . . . 18

2.2 Carte spatiotemporelle obtenue par accumulation de la ligne virtuelle en rouge dans une séquence vidéo . . . . . 18

2.3 Illustration de patchs avec des textures différentes [Ric11] . . . . . 20

2.4 Illustration du problème de l'ouverture : (a) stable, (b) problème classique de l'enseigne du coiffeur, (c) région sans textures. Les deux images $I(0)$ (jaune) et $I(1)$ (rouge) sont superposées. Le vecteur rouge $u$ indique le déplacement du centre du patch bleu [Ric11] . . . . . 21

2.5 Exemple de points d'intérêt extraits avec le détecteur de Harris . . . . . 23

2.6 Illustration de l'hypothèse de constance de l'intensité pour deux images successives [Ric11] . . . . . 26

2.7 Illustration des vecteurs obtenus grâce à l'algorithme de calcul du flux optique . 28

2.8 Représentation d'un vecteur de flux optique dans un repère cartésien . . . . . 29

2.9 Illustration d'un volume spatiotemporel obtenu suite à l'empilement de silhouettes à travers le temps . . . . . 31

2.10 Représentation d'Images d'Historique du Mouvement (MHI), (a) Images clés d'un exercice d'étirement des bras, (b) MHI correspondants aux images clés . . 32

2.11 Représentation des points d'intérêt spatiotemporels de Laptev et Lindeberg [LCSL07] pour des actions similaires exécutées par des personnes différentes . 40

2.12 Illustration des trajectiores extraites par l'approche de Messing et al. [MPK09] 41

2.13 Quatre motifs en (b) de mouvements extraits depuis les trajectoires en (a) en utilisant l'approche de M. Hu et al. [HXF$^+$06] . . . . . 44

2.14 Illustration de l'approche de Wang et al. [WTG06], (a) détection et suivi d'objets pour estimer les trajectoires, (b) motifs de mouvement estimés . . . . . 45

2.15 Approche de W. Hu et al. [HAS08b] appliquée à une vidéo de pèlerinage en (a), (b) champ de vecteurs de mouvement, (c) motifs de mouvement . . . . . 46

2.16 Carte de direction correspondant à une voiture tournant à gauche . . . . . 47

2.17 Illustration du classificateur de Stauffer et Grimson [SG00] . . . . . 48

2.18 Illustration de l'approche d'Andrade et al. [ABF06]. (a) Foule simulée, (b) Données de flux optique simulées . . . . . 50

2.19 Illustration de l'approche d'Adam et al. [SG00]. (a) Une scène typique contenant des moniteurs représentés par des points rouges, (b) la magnitude moyenne du flux optique observée par chaque moniteur . . . . . 52

2.20 Illustration des moyennes des gaussiennes d'un mélange Gaussien à quatre dimensions (x, y, vx, vy). Les lignes rouges représentent la direction et la magnitude moyenne dans la position moyenne, tandis que les ellipses blanches sont proportionnelles aux variances des positions . . . . . 53

2.21 Modèle humain avec caméra verticale . . . . . 56

2.22 Détection des blobs : (a) Soustraction de l'arrière-plan. (b) Remplissage des vides 56

2.23 Résultats de la segmentation dans une image : (a) Blobs détectés. (b) Résultats de la segmentation . . . . . 56

2.24 Segmentation humaine multiple. a) Image d'origine. b) Premier plan. c) Analyse de la bordure dans une région . . . . . 58

2.25 Bords statiques et mobiles d'une image. (a) Carte des bords statiques. (b) Bords mobiles . . . . . 59

2.26 Résultat du regroupement des segments reliés . . . . . 59

2.27 Détection et suivi des personnes sur un quai de métro . . . . . 60

2.28 L'algorithme de correspondance de forme de Zhao. Les images s'affichent dans le sens des aiguilles d'une montre, de la droite vers la gauche : image d'entrée, calcul de la fonction de coût, tête finale des candidats, image de l'orientation des bords et carte étendue de l'image des gradients $O$ . . . . . 61

2.29 Illustration d'un système de comptage stéréoscopique (a) Système de comptage où la caméra stéréo se trouve au-dessus de la porte, (b) Echantillons d'images obtenues avec une caméra stéréo . . . . . 63

2.30 Echantillons d'images lors de l'acquisition. (a) Aucune personne, (b) Personne isolée, (c) et (d) Situations de foule . . . . . 64

2.31 Exemple d'empilement. De droite à gauche : empilements noirs, gradients, et blancs . . . . . 65

3.1 Modèle directionnel pour l'action 'answerPhone'. (a) image courante, (b) vecteurs de flux optique, (c) modèle directionnel associé à la séquence vidéo . . . 71

3.2 Illustration des résultats de notre algorithme de regroupement de données circulaires . . . . . 73

3.3 Échantillon d'images avec les modèles de direction et de magnitude qui leur sont associés . . . . . 75

3.4 Regroupement des blocs dans une image . . . . . 80

4.1 Schéma de l'approche de reconnaissance d'actions . . . . . 85

4.2 Illustration de notre processus de reconnaissance des actions en choisissant la plus petite distance entre le modèle requête et les modèles de référence . . . . . 86

4.3 Résultats de l'ensemble de données KTH. (a) Echantillon d'actions, (b) Matrice de confusion utilisant un bloc de $5 \times 5$ . . . . . 89

4.4 Résultats pour la base ADL. (a) Echantillons d'actions. (b) Matrice de confusion utilisant des blocs de $5 \times 5$ pixels . . . . . 90

4.5 Influence de la taille des blocs et de la combinaison des actions sur la précision 91

5.1 Motifs de mouvement . . . . . 96

5.2 Schéma de notre approche d'extraction des motifs de mouvement . . . . . 98

5.3 Motifs de mouvement à partir d'un modèle directionnel de 3 lignes et 3 colonnes 99

5.4 Motifs de mouvement détectés dans une scène urbaine. (a) Échantillon de la séquence, (b) Motifs de mouvement extraits. Ils sont mieux visibles sur un document en couleurs . . . . . 101

5.5 Motifs de mouvements détectés dans une scène de pèlerinage. (a)Échantillon de la séquence, (b) Motifs de mouvements extraits . . . . . 102

5.6 Les motifs de mouvement détectés par l'approche de Hu et al. [HAS08b] sur la scène de pèlerinage. Chaque couleur indique un motif différent . . . . . 103

5.7 Les motifs de mouvement extraits d'une scène routière complexe . . . . . 103

5.8 Les motifs de mouvement extraits d'une autre séquence de pèlerinage . . . . . 104

5.9 Les motifs de mouvement extraits à l'entrée d'un escalator . . . . . 104

5.10 Schéma de notre approche de détection d'évènements de foule . . . . . 106

5.11 Représentation de groupes en fusion . . . . . 109

5.12 Représentation des évènements dispersion locale (au milieu) et séparation (à droite) . . . . . 110

5.13 Représentation de l'évènement évacuation . . . . . 111

5.14 Les matrices de confusion obtenues en utilisant le classificateur 'Forêt aléatoire', (a) les évènements marche et course, (b) Les évènements fusion, division, dispersion locale et évacuation . . . . . 112

5.15 Comparaions de différentes méthodes de détection d'évènements sous forme de graphique, (a) Précision, (b) Rappel . . . . . 114

5.16 Échantillon de détection d'évènements. Les évènements détectés apparaissent en bleu . . . . . 115

5.17 Exemples de configuration de la caméra : (a) Caméra orientée verticalement au-dessus de la tête des passants, (b) Caméra orientée obliquement . . . . . 117

5.18 Schéma de notre approche d'estimation des flux . . . . . 118

5.19 Étapes clés de notre approche : (a) Sélection de la ligne de comptage, (b) Carte spatiotemporelle des orientations, (c) Détection en ligne des blobs, (d) Représentation de l'orientation des blobs sur la carte spatiotemporelle . . . . . 119

5.20 Résultats du comptage avec caméra oblique . . . . . 122

5.21 Résultats du comptage avec caméra verticale . . . . . 123

5.22 Résultats du comptage dans une séquence vidéo issue de la base PETS2009 . . 123

6.1 Aperçu de notre application. . . . . 132

A.1 Actions de la vie quotidienne, (a) écrire sur un tableau, (b) boire de l'eau . . . . 155

A.2 Évènement de foule correspondant à une évacuation d'urgence (mise en scène fictive) . . . . . 156

A.3 Les vecteurs de flux optique . . . . . 156

A.4 Illustration d'un cercle unitaire où on mesure l'angle $\theta_p$ . . . . . 157

B.1 Les différents partenaires du projet MIAUCE . . . . . 160

# Liste des tableaux

2.1 Comparaison des différentes approches de définition des régions d'intérêt . . . 25

2.2 Comparaison entre les méthodes de description du mouvement par rapport à certains facteurs . . . 33

2.3 Tableau synthétisant les avantages et inconvénients des méthodes de reconnaissance des actions humaines . . . 41

2.4 Tableau synthétisant les avantages et inconvénients des approches d'extraction des motifs de mouvement . . . 49

2.5 Tableau synthétisant les avantages et les inconvénients des approches de détection d'évènements de foule . . . 54

2.6 Tableau synthétisant les avantages et inconvénients des approches d'estimation des flux . . . 66

4.1 Comparaison pour 2 bases de vidéos . . . 91

5.1 Comparaison des résultats de notre approche avec la vérité terrain . . . 104

5.2 Comparaions de différentes méthodes de détection d'évènements, NC signifie que les résultats n'ont pas été communiqués, ND signifie que l'évènement n'est pas détecté par l'approche . . . 114

5.3 Description des ensembles de données . . . 122

5.4 Précision globale du système de comptage . . . 124

# Chapitre 1

# Introduction

## 1.1 Contexte

L'essor des systèmes d'acquisition et de traitement de la vidéo joue un rôle important dans la vie quotidienne. Cette importance évolue proportionnellement avec nos besoins d'automatiser le processus d'extraction de l'information depuis la vidéo.

Le domaine de la vision par ordinateur (appelée aussi vision artificielle, vision numérique ou vision cognitive) a pour but de reproduire sur un ordinateur les capacités d'analyse et d'interprétation propres de la vision humaine. Elle s'inscrit dans le cadre général de la recherche de moyens susceptibles de doter l'ordinateur d'une intelligence comparable à celle des êtres humains (apprentissage, représentation, raisonnement). L'objectif général est de concevoir des modèles et des systèmes capables de représenter et d'interpréter le contenu visuel d'une scène.

La vision par ordinateur est une thématique passionnante de la recherche en intelligence artificielle. Bien que des progrès considérables aient été effectués, la résolution de certains problèmes tels que l'analyse du comportement humain de façon robuste, reste un challenge.

L'analyse du comportement humain s'avère très utile dans un grand nombre d'applications parmi lesquelles :

(i) **La conception et l'agencement des espaces :** l'analyse du comportement humain dans les espaces intérieurs et extérieurs permet d'estimer le taux de fréquentation de chaque espace et la façon dont il a été parcouru. Elle fournit ainsi des lignes directrices afin d'adapter la conception et l'agencement d'un espace selon les habitudes des personnes. Par exemple : l'amélioration de l'agencement des rayons des centres commerciaux, l'optimisation de l'utilisation de l'espace des bureaux ou l'organisation des visites dans un musée. La Figure 1.1(a) montre le taux de fréquentation d'un magasin calculé par notre système d'estimation des flux (présenté dans la Section 5.3).

(ii) **L'interaction homme-machine :** les systèmes de reconnaissance des actions d'une personne permettent de développer des interfaces obéissant aux mouvements d'un utilisateur. Les environnements virtuels et les systèmes de réalité augmentée prenant en compte le comportement des personnes sont plus immersifs. La Figure 1.1(b) illustre le système de

réalité virtuelle VhCVE [BHQ+09] qui utilise la vision par ordinateur pour détecter les expressions faciales et les transposer sur des avatars animés.

(iii) **L'indexation des vidéos :** L'extraction automatique d'informations à partir des vidéos permet d'améliorer leur archivage et leur consultation. Avec l'avènement des sites de partage de vidéos et la démultiplication de la capacité de stockage, il est devenu nécessaire de développer des applications capables d'accéder aux vidéos de façon rapide et efficace. Des connaissances sur les évènements contenus dans les vidéos permettent d'améliorer la pertinence des réponses proposées aux utilisateurs désireux de retrouver du contenu comportant certaines actions humaines.

(iv) **La gestion de la foule :** l'étude automatique du comportement de foule permet de développer des stratégies de gestion de foule, notamment pour des évènements populaires ou fréquentés par un nombre important de personnes (ex. les rencontres sportives, les grands concerts, les manifestations publiques, etc.). L'étude des foules permet d'anticiper certains comportements à risque et d'éviter les accidents causés par le mouvement désorganisé des foules. La Figure 1.1(c) illustre une séquence contenant une forte densité de personnes où les risques d'accident sont élevés, notamment pour la personne soulevée au milieu de la scène.

(v) **La vidéo-surveillance automatique :** elle permet de détecter automatiquement des évènements et apporte des informations variées pour l'assistance des agents de sécurité. A titre d'exemple : le contrôle automatique des entrées et des sorties de certaines zones, l'identification et la reconnaissance des personnes, la détection d'activités inhabituelles, etc. La Figure 1.1(d) illustre un évènement de séparation d'un groupe en deux. Un trait blanc indique les deux groupes concernés.

Nous avons choisi de porter notre attention sur la vidéo-surveillance automatique. Celle-ci repose sur l'utilisation de machines pour assister les opérateurs humains. Dans un premier temps, nous décrivons la vidéo-surveillance automatique, puis nous retraçons son évolution à travers trois générations en soulignant les avancées dans le domaine matériel et logiciel.


(a) Système d'estimation du taux de fréquentation des rayons d'un magasin


(b) Le système de réalité virtuelle VhCVE

(c) Illustration d'une situation inhabituelle dans une scène de foule


(d) Évènement de séparation de groupes détecté par un système de vidéo-surveillance automatique

FIGURE 1.1 – Illustration de quelques applications de l'analyse du comportement humain depuis la vidéo

La vidéo-surveillance automatique est devenue nécessaire de nos jours. L'objectif de la plupart des systèmes de vidéo-surveillance est d'assurer la sécurité, mais ils peuvent être employés pour d'autres fins telles que l'estimation des flux ou l'amélioration de l'agencement des espaces publiques. En effet, en plus d'améliorer l'efficacité de la surveillance dans les lieux publics, ils permettent l'observation des flux, la gestion des ressources, la détection en temps réel des situations critiques et la notification des opérateurs le cas échéant. Les algorithmes de vision par ordinateur permettent d'enrichir les systèmes de vidéo-surveillance d'une certaine intelligence, on parle alors de *vidéo-surveillance automatique*.

Un système de vidéo-surveillance automatique peut être défini comme l'ensemble des outils informatiques et technologiques qui ont pour but d'assister l'opérateur humain. La Figure 1.2 [1] montre un système de vidéo-surveillance automatique. Les écrans sur le mur affichent les flux vidéo bruts alors que les ordinateurs posés sur la table sont utilisés pour assister les opérateurs.

FIGURE 1.2 – Illustration d'un système de vidéo-surveillance automatique

L'objectif principal d'un système de vidéo-surveillance automatique est d'alléger la charge de travail de l'opérateur qui surveille plusieurs moniteurs et d'analyser les flux vidéo en temps réel. Les opérateurs sont ainsi plus réactifs et les erreurs d'inattention sont moins fréquentes. Ceci est notamment utile pour la surveillance des zones publiques, la prévention des incidents, et la collecte de preuves juridiques. Aujourd'hui, avec l'augmentation du nombre de caméras de surveillance, les moyens informatiques utilisés pour l'interprétation des flux vidéo doivent être performants.

Les premiers systèmes de vidéo-surveillance datent des années 60. A l'époque, les caméras analogiques étaient le seul moyen d'acquisition d'images. Les ordinateurs ne permettaient pas encore d'assister les opérateurs à cause des faibles avancées matérielles et logicielles. Puis, sont

1. source : google images

arrivées les caméras numériques permettant de filmer avec une meilleure résolution d'images et réduction du bruit. Dans les années 80, des ordinateurs encore plus puissants ont vu le jour. Cette évolution technologique permet de classer les systèmes de vidéo-surveillance en trois générations [Pet00].

***(a) Première génération de systèmes de vidéo-surveillance (1GSS)***

Les systèmes de vidéo-surveillance de première génération (1GSS, 1960-1980) sont basés sur des sous-systèmes analogiques pour l'acquisition des images, la transmission et le traitement. Ils ont étendu la portée visuelle de l'homme en transmettant les flux vidéo de caméras installées sur un ensemble de sites vers une salle de contrôle centrale. Ces flux s'affichent alors sur des moniteurs.

Ces systèmes ont cependant plusieurs inconvénients et limites car ils nécessitent une bande passante élevée, ce qui complique l'archivage et la récupération des flux vidéo. Le travail des opérateurs humains est aussi difficile car ils ne sont pas assistés par les ordinateurs. Les générations suivantes viennent résoudre ces inconvénients en tirant parti des progrès matériels et logiciels.

***(b) Deuxième génération de systèmes de vidéo-surveillance (2GSS)***

Les systèmes de vidéo-surveillance de deuxième génération (2GSS, 1980-2000) étaient des systèmes hybrides dans le sens où ils ont combiné des sous-systèmes analogiques et numériques pour résoudre certains inconvénients de leurs prédécesseurs. Ils ont exploité les premières avancées des méthodes de traitement de la vidéo numérique en permettant d'avertir les opérateurs de certains évènements. La plupart des travaux de cette période se sont concentrés sur le développement des techniques de détection d'évènements en temps réel pour la vidéo-surveillance automatique.

***(c) Troisième génération de systèmes de vidéo-surveillance (3GSS)***

La troisième génération (3GSS, à partir de 2000) vise à fournir des solutions totalement numériques, du capteur jusqu'à la présentation d'informations symboliques, textuelles, sonores

et visuelles aux opérateurs. Cette génération profite du faible coût du matériel, de la disponibilité élevée des unités de calcul et de la possibilité de communiquer avec des appareils mobiles et des réseaux à faible débit. L'objectif principal de cette génération est d'émettre une alarme en temps réel pour aider les opérateurs à repérer instantanément des évènements.

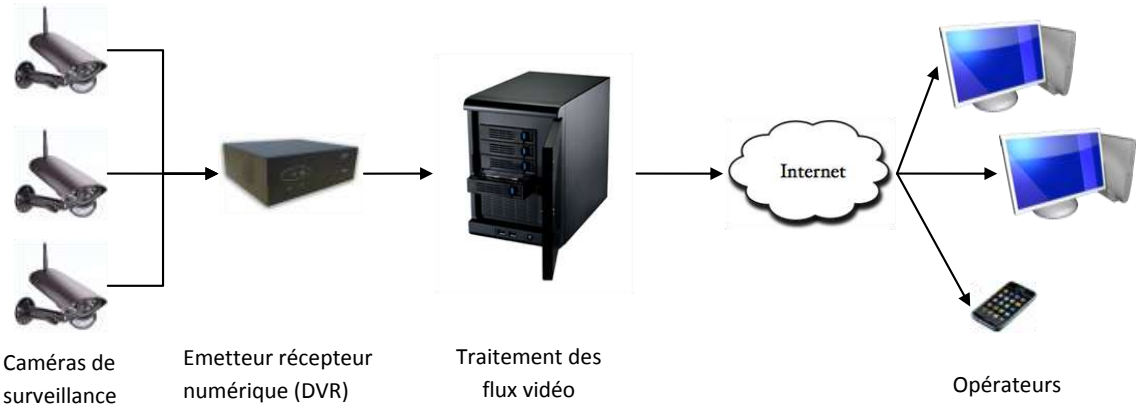



FIGURE 1.3 – Schéma d'un système de vidéo-surveillance de troisième génération typique

La Figure 1.3 dresse une architecture classique d'un 3GSS composé de 4 parties interconnectées. Tout d'abord, les caméras de surveillance envoient les flux vidéo à un dispositif qui permet de numériser les images reçues tout en effectuant des traitements préliminaires comme l'élimination du bruit. Les flux vidéo numérisés sont ensuite envoyés à la couche de traitement des données qui est capable de les analyser en temps réel. Cette couche est considérée comme la partie centrale du système. Ses tâches sont diverses et varient selon les besoins des opérateurs. Ces derniers sont alertés en temps réel grâce à une multitude de moyens de communication. Par exemple, ils peuvent observer un signal visuel particulier sur les moniteurs ou recevoir des messages de notification sur leurs téléphones. Ils peuvent ainsi réagir plus rapidement et efficacement.

La partie *traitement des flux vidéo* est considérée comme le cerveau d'un 3GSS car il a pour but d'analyser le flux vidéo reçu. Parmi les domaines traités par cette partie, on trouve *l'analyse du comportement humain*. Ce domaine pose des problèmes de plus en plus complexes et attire de plus en plus de chercheurs et d'industriels.

L'analyse du comportement humain depuis la vidéo est un domaine riche en expérimentations diverses et variées. Nous utilisons le terme comportement en vision par ordinateur pour dénoter le comportement physique et apparent produit suite à un mouvement, sans interprétation psychologique ou cognitive. Bien que le terme de comportement ne soit pas utilisé dans la majorité des cas, le problème est identique lorsqu'on se réfère à actions, évènements, motifs de mouvement ou estimation des flux. Ce problème consiste à instancier un ensemble de modèles de comportements depuis les flux vidéo. Un modèle sera reconnu lorsqu'il correspond à une certaine interprétation.

Dans le cadre de la vidéo-surveillance automatique, l'analyse du comportement de personnes présentes dans la scène permet de prédire des incidents et de détecter certains évènements. Elle permet également d'obtenir des informations qui expliquent l'activité de ces personnes ainsi que leurs intérêts.

Notre travail de thèse apporte de nouvelles approches pour traiter plusieurs problèmes liés à l'analyse du comportement humain depuis la vidéo. Dans la section suivante, nous mettrons en lumière cette problématique que nous avons abordée durant le travail de thèse.

## 1.2 Problématique

L'analyse du comportement humain depuis la vidéo est un domaine vaste de la vision par ordinateur. Cependant, très peu d'approches ont essayé d'aborder ce problème en ré-exploitant les descripteurs pour traiter les différents problèmes de ce domaine tout en garantissant des résultats fiables et une exécution en temps réel dans le cas échéant.

Nous proposons dans ce travail de développer des descripteurs de niveau intermédiaire basés principalement sur l'orientation et la vitesse du mouvement. Ces descripteurs sont polyvalents et peuvent être réutilisés dans plusieurs problèmes liés à l'analyse du comportement humain. Parmi ces problèmes, nous en faisons ressortir quatre comme faisant partie des objectifs de ce travail de thèse. La section suivante détaille chacun des objectifs.

## 1.3 Objectifs

Selon le nombre de personnes présentes dans la scène, nous sommes amené à analyser des comportements différents. Par exemple, dans une scène de foule, on s'intéresse à l'analyse du comportement de la foule dans sa globalité. Par contre, en présence d'une seule personne, on s'intéresse à analyser l'action effectuée par cette dernière. Notre travail s'articule donc autour de l'analyse du comportement humain dans deux types de scènes :

1. **Analyse du comportement humain dans des scènes individuelles :** Ces scènes se caractérisent par la présence d'une seule personne à la fois dans le champ de la caméra. Nous analysons les vidéos où la personne y effectue une action de la vie courante (ex : marcher, répondre au téléphone, sauter...). Ces actions répondent à des modèles de mouvements réalisés durant un laps de temps court. Ils peuvent être cycliques (ex. l'action marcher) ou non cycliques (ex. l'action ouvrir un livre). Notre objectif est de reconnaitre l'action effectuée par une personne parmi une liste d'actions possibles.
2. **Analyse du comportement humain dans des scènes de foule :** Ces scènes se caractérisent par la présence d'un grand nombre de personnes dans le champ de la caméra. Cependant, il se peut que ces scènes ne contiennent aucune personne. Par exemple dans une voie piétonne, lorsque les piétons sont autorisés à traverser, la voie contiendra un nombre important de personnes qui se déplacent dans des directions différentes. Ceci est parfois appelé mouvement de chaos. Néanmoins, quand les piétons ne sont plus autorisés à traverser, la voie devient vide.

   Nous traiterons trois problèmes dans ce type de scène, à savoir, l'extraction des motifs de mouvement, la détection des évènements de foule et l'estimation des flux. Nous définissons brièvement chaque axe :

   (a) Extraction des motifs de mouvement : consiste à synthétiser les mouvements d'une scène à l'aide d'une carte représentant les tendances de mouvements les plus importantes.

(b) Détection des évènements de foule : consiste à détecter certaines situations liées aux comportements d'un ensemble de personnes.

(c) Estimation des flux : consiste à estimer le nombre de personnes qui traversent une ligne virtuelle ainsi que le sens de déplacement.

Nous proposons une approche unifiée permettant de traiter ces quatre problèmes (reconnaissance des actions, extraction des motifs de mouvement, détection d'évènement de foule et estimation du flux des personnes). Dans ce qui suit, nous décrivons le schéma général de notre approche ainsi que nos contributions dans chacun des quatre problèmes.

## 1.4 Contribution et originalité

Notre travail de thèse traite quatre problèmes de l'analyse du comportement humain. Trois concernent les scènes de foules et un, les scènes individuelles. Nous décrivons les contributions apportées pour chaque problème :

*Analyse du comportement dans des scènes individuelles :* Une méthode de reconnaissance d'actions basée sur la construction de modèles d'orientation et de vitesse de mouvement est proposée. Cette méthode permet la détection des actions effectuées dans des environnements intérieurs et extérieurs. Pour cela, nous définissons deux modèles basés sur le mouvement : le modèle de magnitude et le modèle directionnel.

Le modèle directionnel permet d'estimer les orientations de mouvement les plus fréquentes dans une scène, alors que le modèle de magnitude estime les vitesses de mouvement les plus fréquentes. Une action peut alors être définie par son modèle directionnel et son modèle de magnitude. La reconnaissance de l'action est réalisée en comparant ses modèles à ceux des séquences d'apprentissage.

*Analyse du comportement dans des scènes de foule :* Nous apportons plusieurs contributions dans le domaine de l'analyse des scènes de foule à travers trois axes : l'extraction des motifs de mouvement, la détection des évènements de foule et l'estimation des flux.

- Extraction des motifs de mouvement : une méthodologie est proposée pour estimer les motifs de mouvement. Des parties en mouvement ayant la même orientation sont perçues comme faisant partie du même motif de mouvement. Notre approche s'appuie sur le regroupement des orientations de mouvement dominantes à travers l'estimation du modèle directionnel. Ceci permet de traiter efficacement les scènes où les mouvements sont complexes et non uniformes.
- Détection des évènements de foule : Un ensemble de métriques sur le comportement de la foule est proposé. Parmi ces métriques on trouve le nombre de groupes où un groupe est un ensemble de personnes proches se déplaçant vers la même direction. On définit également comme métrique le degré de dispersion et de regroupement des groupes et leur vitesse moyenne. Les valeurs de ces métriques permettent de détecter des évènements particuliers tels que la fusion, la séparation ou la course.
- Estimation des flux : une approche permettant de compter le nombre de personne traversant une ligne virtuelle (ou ligne de comptage) de façon robuste, rapide et précise est proposée. Elle se base sur le mouvement des personnes traversant la ligne virtuelle. Les paramètres du système sont estimés grâce aux méthodes d'apprentissage ou manuellement. Cela permet à notre système de supporter plusieurs angles de vue et de compter aussi bien les objets rigides (ex. une voiture) que les objets non-rigides (ex. un humain).

En plus des contributions précédentes, nous proposons de traiter ces quatre problèmes à l'aide de nouveaux descripteurs et d'exploiter une approche pyramidale que nous introduisons dans la section suivante.

## 1.5 Schéma général de l'approche

Nous présentons une méthodologie sous forme de pyramide à trois niveaux pour traiter les objectifs de ce travail de thèse. Globalement, notre approche remonte du flux vidéo jusqu'à

l'analyse du comportement humain en passant par trois niveaux tel qu'illustré dans la Figure 1.4.

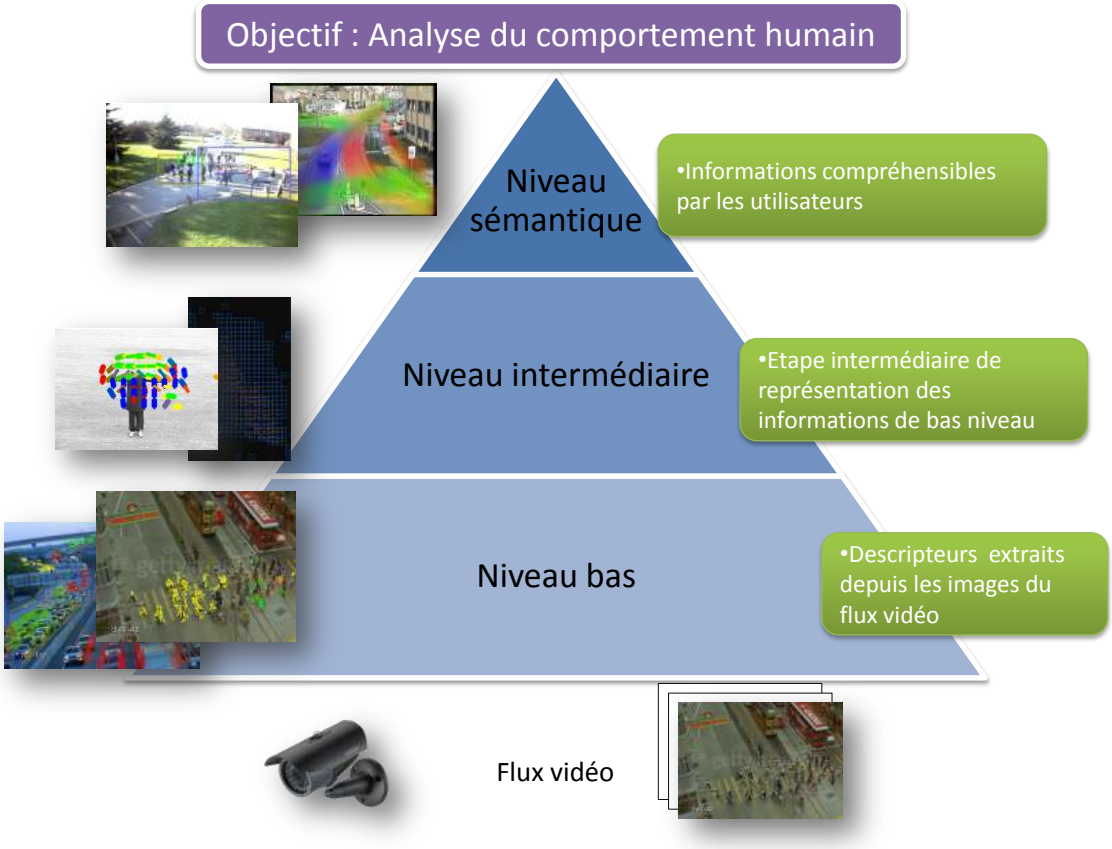



FIGURE 1.4 – Illustration de notre approche globale pour l'analyse du comportement humain

Tout d'abord, le flux vidéo est obtenu à l'aide d'une caméra monoculaire. Ensuite, on applique au flux vidéo un ensemble d'algorithmes qui permettent d'en extraire des informations compréhensibles par les humains. Chaque niveau se sert du niveau précédent tout en augmentant la sémantique.

Le premier niveau de la pyramide dénote le calcul des caractéristiques de *niveau bas*. Ce niveau a pour rôle d'extraire les informations depuis les images d'un flux vidéo grâce aux techniques de traitement d'images ; par exemple les points d'intérêt (ou points caractéristiques) ou les zones en mouvement sont tous deux des caractéristiques de bas niveau. Les caractéristiques

de bas niveau sont génériques et peuvent être réutilisées dans des applications différentes. Toutefois, elles ne portent que très peu d'informations d'ordre sémantique.

Le *niveau intermédiaire* englobe les descripteurs calculés à partir des caractéristiques de bas niveau ; par exemple la trajectoire de déplacement, la vitesse moyenne et la direction de mouvement moyenne. Ces informations peuvent être utiles pour générer des statistiques. Cependant, elles peuvent être très complexes, à haute dimensionnalité et difficiles à interpréter. Elles nécessitent alors d'être traitées de façon plus approfondie dans le niveau suivant. Notons que le niveau intermédiaire ne contient aucune information de bas niveau.

Le *niveau sémantique* dépend entièrement du domaine d'application. Son but est de reconstituer à partir des données du niveau intermédiaire des résultats sur l'analyse du comportement humain qui sont compréhensibles par les utilisateurs. Ce niveau permet par exemple d'aider à la prise de décision dans le cadre d'une stratégie markéting et permet de notifier des opérateurs dans le cadre de la vidéo-surveillance automatique.

## 1.6 Plan de la thèse

Après cette introduction du mémoire de thèse qui fait ressortir notre approche globale et nos objectifs, le Chapitre 2 présente un état de l'art sur l'analyse du comportement humain depuis la vidéo. Nous présentons tout d'abord les approches d'estimation du mouvement les plus célèbres. Ensuite, nous détaillons les méthodes de l'état de l'art qui portent sur l'analyse du comportement humain depuis la vidéo dans les scènes individuelles et les scènes de foule.

Le Chapitre 3 présente nos contributions en terme de descripteurs de niveau intermédiaire. Nous introduisons une approche d'estimation des orientations et vitesses du mouvement dominantes qui nous permettent de construire un modèle de magnitude et un modèle directionnel. Nous présentons également deux méthodes de détection des groupes de personnes, une dans le cadre de la détection d'évènements de foule et l'autre dans le cadre de l'estimation des flux. Les éléments introduits dans ce chapitre sont mis en application dans les deux chapitres suivants.

Dans le Chapitre 4, nous traitons un problème relatif aux scènes individuelles qui est la reconnaissance des actions effectuées par une personne. Elle se base sur l'estimation de modèles d'orientation et de magnitude de mouvement. L'action est reconnue en calculant des distances entre les modèles de référence et les modèles calculés. Nous concluons ce chapitre par les expérimentations incluant une étude comparative montrant les performances de notre approche par rapport aux résultats de l'état de l'art.

Le Chapitre 5 s'articule autour de l'analyse des scènes de foule. Nous traitons trois axes : l'extraction des motifs de mouvement, la détection d'évènements et l'estimation des flux. Nos approches s'appuient sur la théorie du gestaltisme et le modèle d'orientation du mouvement introduit dans le Chapitre 3 pour regrouper les vecteurs de mouvement qui ont une orientation similaire. Chacun des trois axes abordés est accompagné d'une expérimentation en mettant en avant les avantages de notre approche par rapport à l'état de l'art.

Nous concluons ce mémoire de thèse au Chapitre 6 en résumant les contributions scientifiques et les perspectives de recherche.

# Chapitre 2

# État de l’art

## 2.1 Introduction

L'analyse du comportement humain depuis la vidéo attire l'intérêt d'un très grand nombre de chercheurs. Bien que les problèmes soient divers et variés, les plupart des approches de la littérature estiment le mouvement comme une information de bas niveau. Elles définissent ensuite des descripteurs de niveau intermédiaire afin de modéliser le comportement étudié. Dans cette optique, nous organisons ce chapitre en deux parties. La première se veut générale et a pour but de présenter des méthodes de localisation, d'estimation et de classification du mouvement. Tandis que la deuxième partie présente des méthodes spécifiques à l'analyse du comportement humain dans les scènes individuelles et les scènes de foule.

La première partie de cet état de l'art commence avec la Section 2.2. Nous y citons les approches les plus utilisées pour définir les zones d'intérêt. Ce sont les endroits de la scène où le mouvement sera estimé. Une fois ces zones définies, un algorithme d'estimation du mouvement est utilisé. Nous présentons les algorithmes les plus connus dans la Section 2.3, puis les méthodes de classification communément utilisées dans la littérature dans la Section 2.4 afin de conclure cette première partie.

La deuxième partie traite tout d'abord dans la Section 2.5 de l'analyse du comportement humain dans une scène individuelle en s'articulant autour du problème de reconnaissance d'actions humaines. La Section 2.6 dresse un état de l'art sur l'analyse des scènes de foule. Nous concluons ce chapitre avec une synthèse dans la Section 2.7.

## 2.2 Définition des zones d'intérêt

L'analyse du comportement des personnes repose principalement sur l'analyse du mouvement dans la vidéo. Avant de procéder à la détection du mouvement, nous devons d'abord définir les régions où le mouvement est censé se produire. Ces zones sont également appelées les régions d'intérêt. Une région d'intérêt peut être un point, une ligne, une forme aléatoire

ou l'ensemble de l'image. Elle peut être définie manuellement ou estimée automatiquement à l'aide d'algorithmes d'extraction de l'arrière-plan ou par détection de points d'intérêt.

### 2.2.1 Zones prédéfinies

La solution la plus simple pour délimiter la zone d'intérêt est la définition d'une zone de mouvement manuellement. Elle s'avère pratique dans les situations où les zones de mouvement sont préalablement connues et qu'elles ne couvrent pas l'intégralité de la scène. On peut définir une zone de mouvement sous forme d'une ligne (dans le cas de l'estimation des flux) ou sous forme d'un masque (compact ou éparse) définissant les zones de mouvement. Le masque de mouvement peut être utilisé conjointement avec d'autres méthodes de détection des zones de mouvement afin de restreindre les zones de traitement à certaines zones particulières.

Un cas particulier de zones prédéfinies est la ligne virtuelle ou ligne de comptage qui est une zone d'intérêt sous forme d'une ligne. La ligne virtuelle est très utilisée dans les applications d'estimation des flux où l'on cherche à compter le nombre de personnes qui traversent une ligne virtuelle placée à l'entrée d'une salle.

La ligne virtuelle permet de construire une carte spatiotemporelle 2D ; la première dimension étant une position sur la ligne virtuelle et la deuxième est le temps. Les pixels se trouvant sur cette ligne forment une colonne dans la carte spatiotemporelle. Les objets passant sur la ligne virtuelle forment des zones spatialement et temporellement connexes qui s'empilent image par image comme illustré dans la Figure 2.1.

La carte spatiotemporelle colorimétrique ainsi obtenue représente les variations des couleurs des pixels sur la ligne virtuelle à travers le temps. Cette étape est encore illustrée par la Figure 2.2 où la carte spatiotemporelle est obtenue par l'accumulation de la ligne virtuelle (représentée en rouge) sur une vidéo de la base PETS2009 [2].

2. http ://www.cvg.rdg.ac.uk/PETS2009/index.html

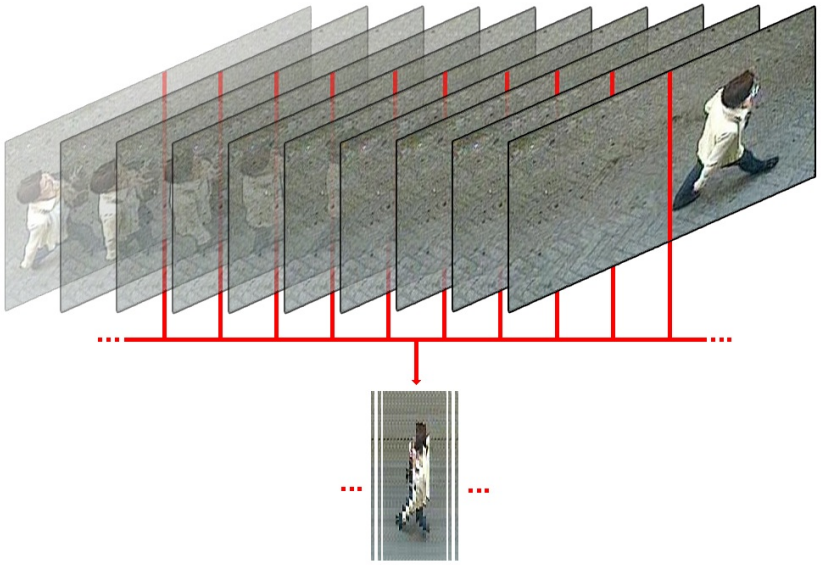


FIGURE 2.1 – Accumulation de la ligne virtuelle à travers le temps

FIGURE 2.2 – Carte spatiotemporelle obtenue par accumulation de la ligne virtuelle en rouge dans une séquence vidéo

### 2.2.2 Avant-plan extrait par une méthode d'extraction d'arrière-plan

La modélisation de l'arrière-plan est utilisée dans différentes applications telles que la vidéosurveillance [CK05, TKBM99] ou le multimédia [EBBV07, PVM07]. La manière la plus simple de séparer l'avant de l'arrière-plan consiste à acquérir une image de l'arrière-plan qui ne contient aucun objet en mouvement. Une soustraction d'images par rapport à un seuil est ensuite

effectuée entre chaque nouvelle image acquise et l'image initiale de l'arrière-plan. Cependant une image unique de l'arrière-plan est souvent indisponible car il est modifié en continu par divers évènements (ex. changements d'éclairage, rajout d'objets, etc.). Pour palier aux problèmes de robustesse et d'adaptation, de nombreux travaux relatifs à la modélisation de l'arrière-plan ont été proposés et peuvent être trouvés dans plusieurs études [BEBV08, Pic04, EESA08].

Les méthodes peuvent être classées dans les catégories suivantes : Modélisation basique de l'arrière-plan [LH02], Modélisation statistique de l'arrière-plan [WADP97b], Modélisation floue de l'arrière-plan [SMP08] et estimation de l'arrière-plan [MMSZ05]. Une autre classification est disponible en termes de prédiction [WS06], de récursivité [CK05], d'adaptation [Por03], ou de modalité [PT05]. Toutes ces méthodes basées sur la modélisation de l'arrière-plan suivent les étapes suivantes : modélisation de l'arrière-plan, initialisation de l'arrière-plan, mise à jour de l'arrière-plan et détection de l'avant-plan.

Dans le contexte d'une application de vidéo-surveillance du trafic routier, Friedman et Russell [FR97] ont proposé de modéliser chaque pixel à l'aide d'un mélange (mixture) de trois Gaussiennes qui correspondent à la route, les véhicules et les ombres. Ce modèle est initialisé au moyen d'un algorithme d'Espérance-Maximisation (EM) [DLR77]. Les Gaussiennes sont ensuite marquées de manière heuristique comme suit : le composant le plus sombre est marqué comme ombre alors que les deux autres composants sont différenciés grâce à la variance (la variance la plus large est étiquetée en tant que véhicule et l'autre en tant que route). La classification de l'avant-plan se fait alors en associant chaque pixel à la gaussienne qui lui correspond. Un algorithme d'EM incrémental est utilisé pour la maintenance de l'arrière-plan afin d'effectuer le traitement en temps réel. Cependant, ce processus souffre d'un manque d'adaptation aux changements qui apparaissent dans la scène à travers temps. Stauffer et Grimson [SG99] ont généralisé cette idée par la modélisation de l'intensité de couleur de chaque pixel sur l'intervalle de temps précédent $\{X_1, ..., X_t\}$ par un mélange de $K$ Gaussiennes.

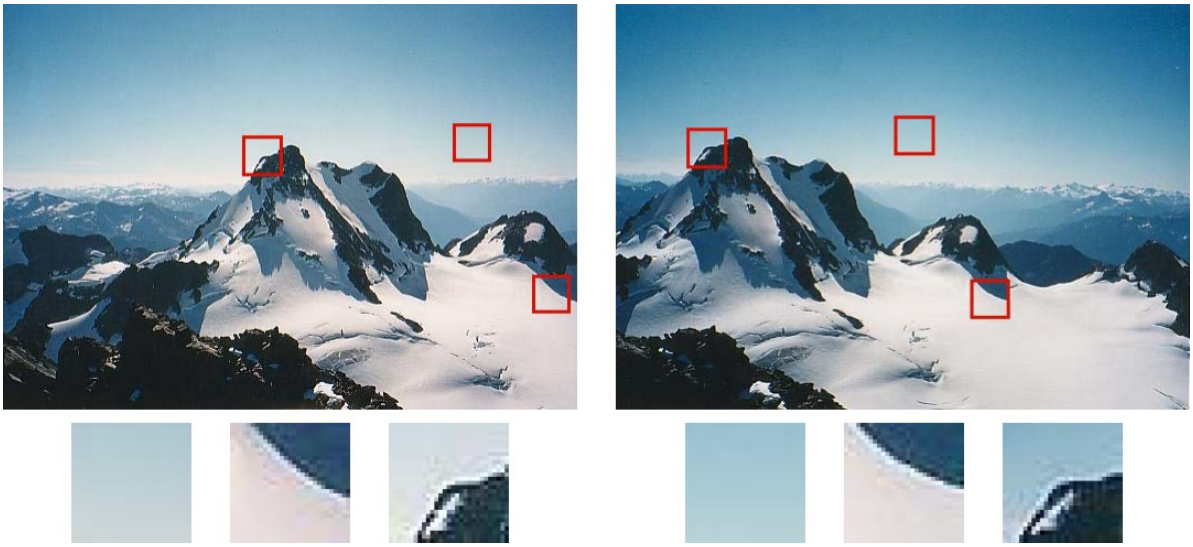
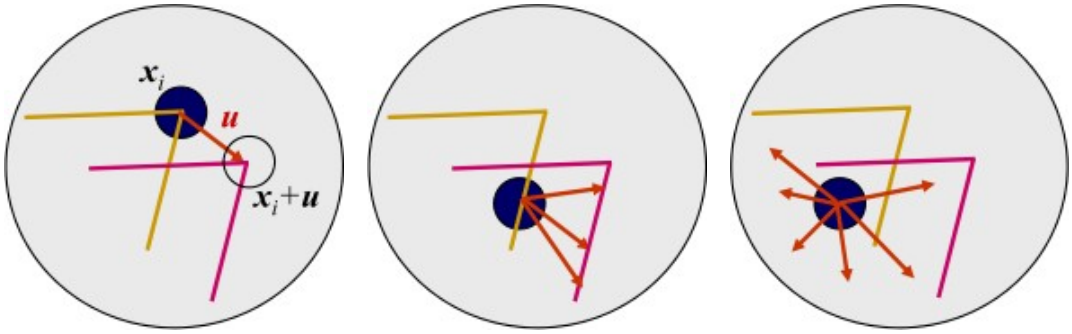


FIGURE 2.3 – Illustration de patchs avec des textures différentes [Ric11]

### 2.2.3 Points d'intérêt

Le terme *point d'intérêt* (en anglais : Point Of Interest ou POI) est issu du domaine de la topologie. Dans ce domaine, il représente un endroit utile ou potentiellement intéressant (selon un critère donné) qui peut être un hôtel, un arrêt de train ou tout autre endroit utile. Pour notre travail de thèse, les points d'intérêts sont des pixels de l'image qui sont susceptibles d'être suivis efficacement dans les images suivantes. Ils sont parfois appelés dans la terminologie anglaise '*good features to track*' (descripteurs qui peuvent être facilement suivis) [ST94, Tri04]. La Figure 2.3 montre trois patchs dans deux images successives. On remarque que les patchs dépourvus de textures sont difficiles à localiser dans l'image suivante (voir Figure 2.4(c)). A contrario, les patchs avec des contrastes (gradients) élevés sont facilement suivis. Cependant, les lignes droites sur une seule orientation souffrent du problème de l'ouverture [HS81b, LK81, Ana89]. Ce problème rend l'alignement des patchs possible uniquement sur une direction normale à la direction du bord (Figure 2.4(b)). Les patchs avec des gradients dans au moins deux orientations différentes sont les plus faciles à suivre, comme illustré dans la Figure 2.4(a).

Nous avons choisi d'utiliser le détecteur de coins de Harris [HS88] pour localiser les points d'intérêt. L'algorithme d'extraction des points d'intérêt de Harris est réputé pour son invariance




(a) (b) (c)

FIGURE 2.4 – Illustration du problème de l'ouverture : (a) stable, (b) problème classique de l'enseigne du coiffeur, (c) région sans textures. Les deux images $I(0)$ (jaune) et $I(1)$ (rouge) sont superposées. Le vecteur rouge $u$ indique le déplacement du centre du patch bleu [Ric11]

à la rotation, au changement d'échelle, à la variation de luminosité et aux bruits [3] dans les images [SMB00]. L'algorithme est rapide, ce qui convient aux applications temps réel. Il est aussi déterministe dans le sens où il retourne toujours les mêmes points d'intérêt pour une image donnée en gardant les mêmes paramètres pour l'algorithme.

Le détecteur de Harris utilise une fonction d'auto-corrélation du signal, où la fonction d'auto-corrélation mesure les changements locaux du signal en déplaçant des patchs par petits intervalles dans diverses directions.

Soit un point $(x,y)$ de déplacement (Δx, Δy), la fonction d'auto-corrélation locale est définie comme :

$$c(x,y) = \sum_{(x_i,y_i)\in w} [I(x_i,y_i) - I(x_i+\Delta x, y_i+\Delta y)]^2 \tag{2.1}$$

où $I(x_i,y_i)$ est l'intensité du pixel à la position $(x_i,y_i)$ de l'image $I$ et les $(x_i,y_i)$ sont les points de la fenêtre $w$ centrée sur $(x,y)$. L'image déplacée est approximée par un polynôme de Taylor d'ordre 1 :

$$I(x_i+\Delta x, y_i+\Delta y) \approx I(x_i,y_i) + [I_x(x_i,y_i)I_y(x_i,y_i)] \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \tag{2.2}$$

3. le bruit est toute information parasite qui peut survenir lors de la phase d'acquisition des images

où $I_x(.,.)$ et $I_y(.,.)$ sont les dérivées partielles sur $x$ et $y$ respectivement. En substituant la partie gauche de l'Equation 2.2 dans l'Equation 2.1 :

$$c(x,y) = \sum_w ([I(x_i,y_i)I_y(x_i,y_i)] \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix})^2 \tag{2.3}$$

$$= [\Delta x, \Delta y] M(x,y) \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \tag{2.4}$$

où :

$$M(x,y) = \begin{pmatrix} \sum_w [I_x(x_i,y_i)]^2 & \sum_w I_x(x_i,y_i)I_y(x_i,y_i) \\ \sum_w I_x(x_i,y_i)I_y(x_iy_i) & \sum_w [I_y\ (x_i,y_i)]^2 \end{pmatrix} \tag{2.5}$$

La matrice $2 \times 2$ symétrique $M(x,y)$ correspond à l'intensité du voisinage local. Soient $\lambda_1$ et $\lambda_2$ les valeurs propres de la matrice $M(x,y)$. Les valeurs propres forment une description invariante à la rotation. Il y a trois cas à considérer :

1. $\lambda_1$ et $\lambda_2$ sont élevées, correspondant à une fonction d'auto-corrélation locale sous forme de pic, et un déplacement dans n'importe quelle direction résultant en une augmentation importante de $c(x,y)$. Ce cas indique un *coin* (Figure 2.4(a)).
2. Une valeur propre élevée alors que la deuxième est faible, la fonction d'auto-corrélation locale a alors une allure qui change peu en $c(x,y)$ dans le sens du bord et change considérablement dans la direction orthogonale. Ce cas indique la présence d'un *bord* (Figure 2.4(b)).
3. $\lambda_1$ et $\lambda_2$ sont petites, $c(x,y)$ subit des petits changements quelle que soit la direction. Le voisinage local est alors d'intensité approximativement constante (Figure 2.4(c)). Ce cas indique la présence d'un coin.

Comme le présente la Figure 2.4(a), les coins sont les points les plus faciles à suivre (d'où le nom *good features to track*). Ces points sont donc caractérisés par des valeurs de $\lambda_1$ et $\lambda_2$ élevées.

FIGURE 2.5 – Exemple de points d'intérêt extraits avec le détecteur de Harris

La figure 2.5 montre un exemple des points d'intérêt de Harris détectés sur image représentant une personne en train de saisir un téléphone. Ce détecteur nous permet ainsi d'avoir des pixels qui seront suivis plus facilement dans les images suivantes. Le suivi est effectué grâce aux algorithmes de flux optique que nous abordons dans la section 2.3.1.

### 2.2.4 Synthèse

Nous avons vu dans cette section différentes approches de définition de régions d'intérêt. Elles peuvent être sous forme de zones prédéfinies, ou de l'avant-plan extrait par une méthode d'extraction de l'arrière-plan ou, encore, sous forme de points d'intérêt. Le choix de la région

d'intérêt dépend de l'application et du type de scène analysée. Il est également possible de combiner ces différentes approches. Par exemple, on peut détecter les points d'intérêt qui sont présents dans une zone prédéfinie.

Le Tableau 2.1 résume les principaux avantages et inconvénients des approches vues précédemment par rapport aux facteurs suivants :

1. Temps d'exécution : le temps que prend une approche pour générer une zone d'intérêt pour une image donnée. Les zones manuellement prédéfinies ne nécessitant aucun traitement informatique, ont un temps d'exécution nul.
2. Adéquation aux scènes individuelles : une scène individuelle contient une seule personne. Une méthode d'extraction de l'arrière-plan extrait la silhouette de la personne qui nous permet d'estimer les positions de certains membres du corps (mains, tête, etc.).
3. Adéquation aux scènes de foule : une scène de foule contient un nombre variable de personnes. Les personnes sont souvent occultées et il est difficile de distinguer leurs silhouettes. Il n'y a donc pas d'intérêt à utiliser une méthode d'extraction de l'arrière-plan dont l'utilité principale est la détection des personnes. Les points d'intérêts sont plus pertinents car ils ne dépendent pas des occlusions.

Nous voyons que les points d'intérêt sont les plus appropriés pour leur rapidité d'exécution et leur efficacité dans les scènes où le nombre de personnes est important. Néanmoins, dans le cadre de l'estimation des flux, il est plus intéressant de définir une zone sous forme de ligne qu'on a appelé ligne de comptage. On peut alors compter le nombre de personnes qui franchissent cette ligne, et ce, sans se soucier du mouvement dans le reste de la scène.

| Facteur | Zones prédéfinies | Extraction de l'arrière-plan | Points d'intérêt |
|---|---|---|---|
| Temps d'exécution | Aucun car elles sont définies au préalable par l'utilisateur | Dépend de la résolution de l'image | Généralement plus rapides que les méthodes d'extraction de l'arrière-plan |
| Scènes individuelles | N'est pas utile si la personne se déplace dans toute la scène | Permet d'extraire la silhouette d'une personne facilitant grandement sa détection et son analyse | Les points détectés sont optimisés pour une bonne estimation du mouvement |
| Scènes de foule | Permet de définir une ligne de comptage | Les occlusions ne permettent pas de tirer partie des silhouettes détectées | Les points détectés sont optimisés pour une bonne estimation du mouvement et ne dépensent pas du nombre de personnes dans la scène |

TABLE 2.1 – Comparaison des différentes approches de définition des régions d'intérêt

## 2.3 Descripteurs de mouvement

Une fois les zones de mouvement détectées, l'étape suivante consiste à quantifier le mouvement. Nous montrons quelques méthodes d'estimation du mouvement : (i) Le flux optique, (ii) les volumes spatiotemporels et (iii) les *images historique et énergie du mouvement*. Nous résumons ensuite les avantages et les inconvénients de ces méthodes dans une synthèse.

### 2.3.1 Flux optique

Le flux optique (ou flot optique) est souvent employé pour représenter le mouvement dans les vidéos. Il permet d'estimer le déplacement des pixels entre deux images successives. La projection d'un mouvement en 3 dimensions vers un plan en 2 dimensions implique que l'estimation du flux optique est un problème mal formé [BB95] (i.e. problème qui n'a pas une solution unique). Néanmoins, certaines méthodes permettent d'estimer efficacement le flux optique.

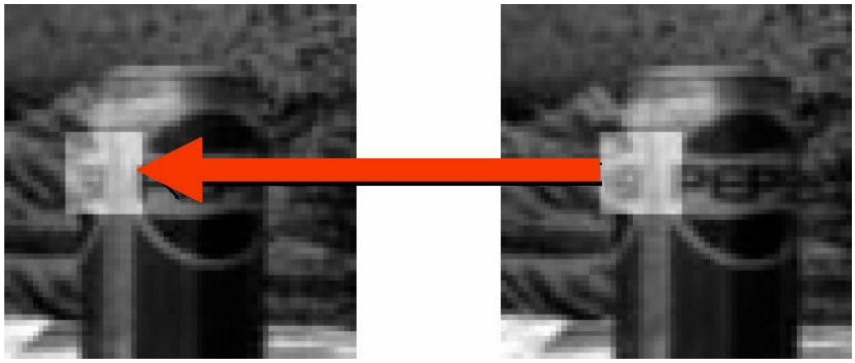


FIGURE 2.6 – Illustration de l'hypothèse de constance de l'intensité pour deux images successives [Ric11]

L'hypothèse initiale pour le calcul du flux optique est que l'intensité des points de l'image à travers le temps est approximativement constante pour des petites durées [HS81a] (voir la Figure 2.6). Formellement si $J(x,y,t)$ est la fonction qui donne l'intensité du point à la position $(x,y)$ de l'image à l'instant $t$, on a :

$$J(x,y,t) \approx J(x+\Delta x, y+\Delta y, t+\Delta t) \tag{2.6}$$

où $\Delta x$ est le déplacement dans l'espace du point $(x,y,t)$ après un temps $\Delta t$. L'écriture sous forme de série de Taylor donne :

$$J(x,y,t) \approx J(x,y,t) + \nabla J.\Delta x + \Delta t J_t + O^2 \tag{2.7}$$

où $\nabla J = (J_x, J_y)$ et $J_t$ sont les dérivées partielles de $1^{er}$ ordre de $J(x,y,t)$. $O^2$ dénote les termes d'ordre 2 et plus, qui sont négligeables et qu'on ignore par la suite. On soustrait $I(x,t)$ et on divise par $\Delta t$ :

$$\nabla J.\vartheta + J_t = 0 \tag{2.8}$$

où $\nabla J = (J_x, J_y)$ est le gradient spatial et $\vartheta = (u,v)$ est la vélocité de l'image. L'équation 2.8 est connue sous le nom de *l'équation des contraintes du flux optique* et définit une contrainte locale unique sur le mouvement de l'image. Cependant, cette contrainte n'est pas suffisante pour

calculer les composantes de $v$ car cette contrainte est un problème mal formé. Cependant, plusieurs méthodes existent pour estimer les vecteurs de flux optique en partant de ces hypothèses.

Nous avons utilisé l'implémentation proposée par Birchfield [Bir96] de l'algorithme de calcul du flux optique KLT [LK81]. Cet algorithme nécessite comme paramètres les pixels de la première image dont on souhaite estimer le déplacement. Ces pixels sont généralement estimés grâce à un algorithme de détection des points d'intérêt (cf. Section 2.2.3).

Comme décrit par Baker et Matthews [BM04], l'algorithme trouve pour chaque point sur la première image, son correspondant sur la deuxième image qui minimise l'équation suivante :

$$\sum_{x,y}[T(x,y)-I(W(x,y;p))]^2 \tag{2.9}$$

où $T$ est l'apparence du point dont on cherche la correspondance dans la deuxième image, $I$ est la première image, un point $(x,y)$ appartient à la fenêtre de recherche de correspondance, $W$ est l'ensemble des transformations envisagées (dans ce cas, la translation) entre la première et la deuxième image et $p$ représente l'ensemble des paramètres de la transformation. Pour calculer les équations de mise à jour du suivi, nous exprimons d'abord l'équation 2.9 comme un problème de mise à jour itérative (en remplaçant $p$ avec $p+\Delta p$), et l'écrivons sous forme de série de Taylor :

$$\sum_{x,y}[T(x,y)-I(W(x,y;p+\Delta p))]^2 \approx \sum_{x,y}[T(x,y)-I(W(x,y;p))-\Delta I\frac{\Delta W}{\Delta p}\Delta p]^2 \tag{2.10}$$

Maintenant, nous prenons la dérivée de l'équation 2.10 par rapport aux paramètres de mise à jour :

$$\begin{aligned}&\tfrac{\Delta}{\Delta p}\textstyle\sum_{x,y}[T(x,y)-I(W(x,y;p))-\Delta I\frac{\Delta W}{\Delta p}\Delta p]^2\\ &=\textstyle\sum_{x,y}[\Delta I\frac{\Delta W}{\Delta p}]^T[T(x,y)-I(W(x,y;p))-\Delta I\frac{\Delta W}{\Delta p}\Delta p]^2\end{aligned} \tag{2.11}$$

En mettant la dérivée de l'équation 2.11 à zéro et en résolvant pour le paramètre de mise à jour $\Delta P$, on obtient :

$$\Delta p = [\sum_{x,y} S^T(x,y)S(x,y)]^{-1}.\sum_{x,y} S^T(x,y)T(x,y) - I(W(x,y;p)) \tag{2.12}$$

où $S(x,y) = \Delta I \frac{\Delta W}{\Delta p}$ est l'image calculée à l'aide de la méthode des gradients [BM04]. En mettant à jour de façon itérative les paramètres selon l'équation 2.12, nous pouvons rapidement trouver le meilleur point correspondant au point suivi.

Le résultat de cet algorithme est un ensemble de vecteurs $V$ illustrés dans la Figure 2.7. Les vecteurs sont colorés selon leur orientation et leur longueur représente la vitesse de mouvement. L'ensemble $V$ est défini comme suit :

$$V = \{V_1 ... V_N | V_i = (X_i, Y_i, A_i, M_i)\}$$

où :

– $X_i$ et $Y_i$ sont les coordonnées de l'origine du vecteur $V_i$.

– $A_i$ est l'orientation du vecteur de mouvement $V_i$.

– $M_i$ est la magnitude du vecteur de mouvement $V_i$.

FIGURE 2.7 – Illustration des vecteurs obtenus grâce à l'algorithme de calcul du flux optique

La Figure 2.8 montre un vecteur de flux optique dans un repère cartésien. Le point $P(X_i, Y_i)$ est la position du point d'intérêt $i$ à l'image $t$, tandis que $Q(X_i, Y_i)$ est sa position dans l'image $t+1$. On peut facilement calculer la distance entre ces deux points grâce à la distance euclidienne :

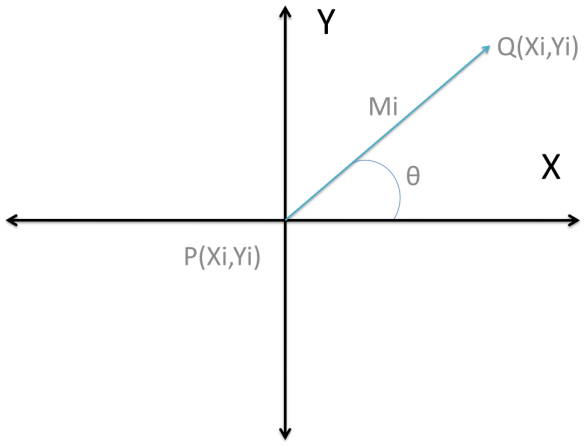



FIGURE 2.8 – Représentation d'un vecteur de flux optique dans un repère cartésien

$$M_i = \sqrt{(Qx_i - Px_i)^2 + (Qy_i - Py_i)^2} \tag{2.13}$$

L'orientation du mouvement $A_i$ du point d'intérêt $i$ est calculée avec la fonction trigonométrique suivante :

$$A_i = atan(\frac{y_i}{x_i}) \tag{2.14}$$

où $x_i = Qx_i Px_i$ et $y_i = Qy_i Py_i$. Cependant, quelques problèmes surviennent si nous souhaitons calculer l'orientation du mouvement à l'aide de l'équation 2.13 ; par exemple :

1. L'équation 2.14 n'est pas valide pour tous les angles compris entre 0 et $2\pi$ car elle ne fait pas la différence entre les angles diamétralement opposés. Prenons par exemple deux points ($x_1 = 1$, $y_1 = 1$) et ($x_2 = -1$, $y_2 = -1$). Avec l'équation 2.14, le point ($x_2 = -1$, $y_2 = -1$) produit le même angle que ($x_1 = 1$, $y_1 = 1$) alors qu'ils sont dans des quadrants différents dans le cercle unitaire.

2. Les points sur l'axe vertical ont $x_i = 0$, or, si nous voulons calculer $(\frac{y_i}{x_i})$ nous aurons $\infty$ qui correspond à un problème de division par zéro.

Afin d'éviter ces problèmes, nous utilisons la fonction $atan2(y_i, x_i)$ qui prend $x_i$ et $y_i$ comme arguments. Désormais, l'orientation de mouvement $A_i$ du point d'intérêt $i$ est calculée de façon précise à l'aide de l'équation suivante :

$$atan2(y_i, x_i) = \begin{cases} \phi sgn(y_i) & si\ x_i > 0, \quad y_i \neq 0 \\ 0 & six_i > 0, \quad y_i = 0 \\ \frac{\pi}{2} sgn(y_i) & si\ x_i = 0, \quad y_i \neq 0 \\ \text{non défini} & si\ x_i\ = 0, \quad y_i = 0 \\ (\pi - \phi) sgn(y_i) & si\ x_i < 0, \quad y_i \neq 0 \\ \pi & si\ x_i < 0, \quad y_i = 0 \end{cases} \tag{2.15}$$

où $\phi$ est un angle dans l'intervalle $[0, \pi/2[$ tel que $tan(\phi) = |\frac{x}{y}|$ et $sgn$ est la fonction signe définie de la façon suivante :

$$sgn(y_i) = \begin{cases} -1 & si\ y_i < 0 \\ 0 & si\ y_i = 0 \\ 1 & si\ y_i > 0 \end{cases} \tag{2.16}$$

La fonction *atan2* gère convenablement les pentes infinies et place l'angle dans le bon quadrant. Par exemple $atan2(1,1) = \pi/4$ et $atan2(-1,-1) = -3\pi/4$.

### 2.3.2 Volumes spatiotemporels

Un volume spatiotemporel (STV ou Spatio-Temporal Volume) 3D est composé d'images empilées d'une séquence vidéo. La dimension temporelle apporte la troisième dimension. Afin de construire ce volume, il est nécessaire de procéder à une soustraction précise de l'arrière-plan.

Blank et al. [BGS+05, GBS+07] empilent d'abord les silhouettes à partir d'une séquence donnée pour former un STV (voir Figure 2.9). Puis ils appliquent l'équation de Poisson pour déduire les caractéristiques de saillance et d'orientation dans l'espace spatiotemporel. Des descripteurs globaux sont ainsi obtenus pour un intervalle temporel donné en additionnant les moments relatifs aux descripteurs locaux. Pour régler les problèmes liés aux longueurs variables des séquences vidéo, Achard et al. [AQMM08] utilisent des sous-ensembles de volumes spatiotemporels pour chaque séquence. Chaque sous-ensemble porte uniquement sur une partie de la dimension temporelle.

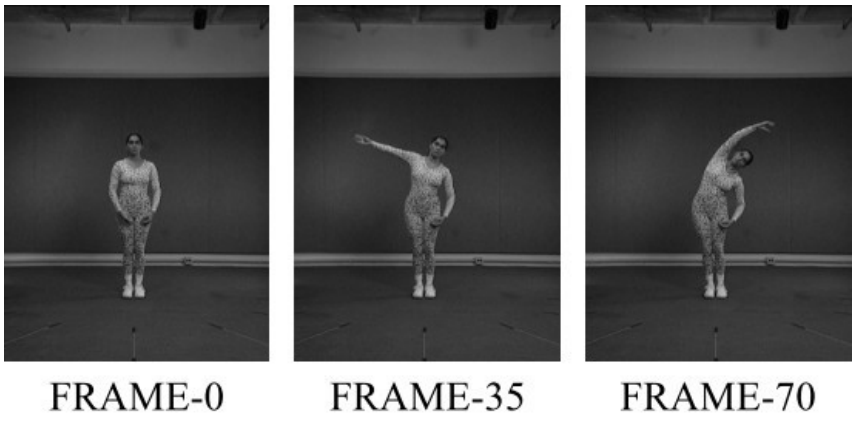

FIGURE 2.9 – Illustration d'un volume spatiotemporel obtenu suite à l'empilement de silhouettes à travers le temps

### 2.3.3 Les images d'historique et d'énergie du mouvement

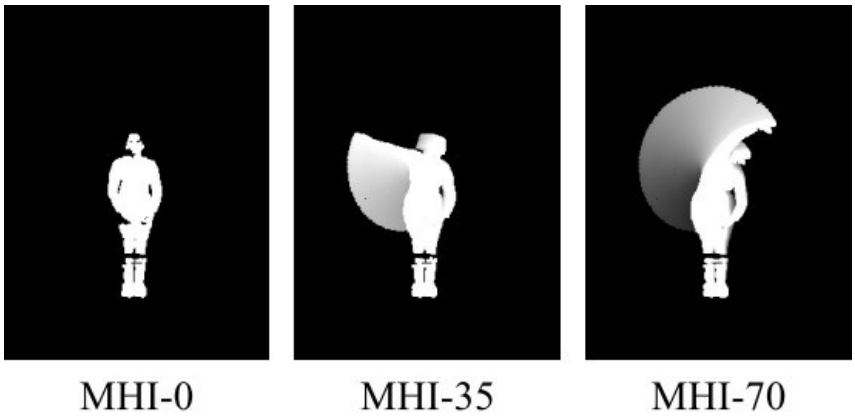
Comme évoqué plus haut, la silhouette d'une personne peut être obtenue en faisant une soustraction de l'arrière-plan. En général, la silhouette comporte du bruit en raison d'une extraction imparfaite. De plus, elle est plus ou moins sensible aux différents angles et encode implicitement l'anthropométrie de la personne. Cependant, elle est porteuse d'un grand nombre d'informations, comme la largeur, la hauteur, la position, la forme, etc. Une fois la silhouette obtenue, il existe de nombreuses façons de modéliser les contours ou la zone dans laquelle elle se trouve.

L'une des utilisations les plus anciennes de la silhouette remonte à Bobick et Davis [BD01]. Les auteurs procèdent à l'extraction des silhouettes à partir d'une seule vue et regroupent les

différences entre les images d'une séquence représentant une action. Cela aboutit à une MEI (Motion Energy Image - Image d'Énergie du Mouvement) dans laquelle se produit le mouvement et une MHI (Motion History Image - Image d'Historique du Mouvement) dont l'intensité des pixels représente une trace des mouvements de la silhouette. La Figure 2.10 illustre des MHI construites à l'aide d'une fonction appliquée à l'historique du mouvement ; plus les pixels sont claires, plus le mouvement est récent. Un inconvénient inné à cette méthodologie est l'auto-occultation du mouvement lorsque des mouvements sont effectués sur la même zone.




(a)




(b)

FIGURE 2.10 – Représentation d'Images d'Historique du Mouvement (MHI), (a) Images clés d'un exercice d'étirement des bras, (b) MHI correspondants aux images clés

### 2.3.4 Synthèse

Cette section a présenté quelques méthodes d'estimation du mouvement qui sont : le flux optique, les volumes spatiotemporels et les images d'historique et énergie du mouvement. Nous les comparons par rapport aux facteurs suivants :

1. Environnement interne : un environnement interne est maitrisé par les utilisateurs et son comportement est connu et prévisible. Il n'y a pas de réel vainqueur entre le flux optique et l'extraction de l'arrière-plan.
2. Environnement externe : à l'opposé d'un environnement interne, un environnement externe n'est pas maitrisable. Les méthodes d'estimation du flux optique sont les plus fiables contrairement aux deux autres méthodes qui reposent sur l'extraction de l'arrière-plan.
3. Scène individuelle : une scène individuelle contient une seule personne. Le flux optique extrait le mouvement de la personne tandis qu'une méthode d'extraction de l'arrière-plan permet d'obtenir sa silhouette. La silhouette nous permet d'estimer les positions de certains membres du corps (mains, tête, etc.).
4. Scène de foule : une scène de foule contient un nombre variable de personnes. Les personnes sont souvent occultées et il est difficile de distinguer leurs silhouettes. Il n'est donc pas intéressant d'utiliser une méthode d'extraction de l'arrière-plan dont l'utilité principale est la détection des personnes. Le flux optique s'avère très utile dans ce genre de scènes car il permet de détecter des groupes qui se déplacent dans la même direction et d'extraire les motifs de mouvement.

| Facteur | Flux optique | Volumes spatiotemporels | MHI |
|---|---|---|---|
| Environnement interne | ✓ | ✓ | ✓ |
| Environnement externe | ✓ | ✗ | ✗ |
| Scène individuelle | ✓ | ✓ | ✓ |
| Scène de foule | ✓ | ✗ | ✗ |

TABLE 2.2 – Comparaison entre les méthodes de description du mouvement par rapport à certains facteurs

Le Tableau 2.2 synthétise la comparaison effectuée plus haut. Nous y voyons clairement que le flux optique est la méthode la plus adaptée pour nos objectifs. Pur cela, dans nos approches, les vecteurs du flux optique extraits depuis la vidéo sont les descripteurs de bas niveau les plus appropriés.

## 2.4 Méthodes de classification basée sur l'apprentissage automatique

La méthode d'apprentissage génère une fonction qui fait correspondre à une image reçue en entrée une étiquette. Les caractéristiques jouent un rôle important dans la classification. C'est pour cela qu'il faut choisir celles qui peuvent discriminer au mieux une classe d'une autre. Différents types de caractéristiques peuvent être utilisés telles que : la couleur, la bordure, la texture, le flux optique, etc. Différentes méthodes d'apprentissage sont disponibles telles que les arbres de décision [GK95], les réseaux de neurones [RBK96], *AdaBoost* (Adaptive Boosting) [VJS03], ou les machines à vecteurs de support (SVM) [POP98].

Certaines méthodes utilisent l'apprentissage semi-supervisé pour réduire la quantité de données qu'il faut annoter manuellement lors de la collecte des données. L'idée est de former deux classificateurs à l'aide d'un petit ensemble de données étiquetées. A la fin de la phase d'apprentissage, chaque classificateur est utilisé pour assigner des données non étiquetées pour l'échantillon d'entrainement de l'autre. Cette méthode a été utilisée avec succès pour réduire la quantité de l'interaction manuelle nécessaire pour l'entrainement en utilisant *AdaBoost* [LVF03] et SVM [KLS03]. Ces méthodes sont présentées par la suite.

### 2.4.1 Les Machines à Vecteurs de Support (SVM)

Les SVM sont utilisées pour regrouper des données en plusieurs classes à l'aide des marges maximales de l'hyperplan qui séparent une classe de l'autre [BGV92]. La marge de l'hyperplan, qui est maximisée, est définie par la distance entre l'hyperplan et les points de données

les plus proches. Les points de données qui se trouvent à la limite de la marge de l'hyperplan sont appelés vecteurs. Pour la détections de personnes, deux classes sont utilisées : la classe *personne* (les échantillons positifs) et classe *pas une personne* (échantillons négatifs). A partir d'exemples d'entrainement annotés manuellement suivant ces deux classes, le calcul de l'hyperplan parmi une infinité d'hyperplans possibles est effectué. SVM a été utilisé par Papageorgiou et al. [POP98] pour la détection de piétons et de visages dans les images.

### 2.4.2 AdaBoost

AdaBoost est un méta-algorithme qui peut être utilisé conjointement avec de nombreux autres algorithmes d'apprentissage afin d'améliorer leurs performances. Le boosting est une méthode itérative qui trouve une classification très précise, en combinant un grand nombre de classificateurs de base [FS95]. Dans la phase d'entrainement de l'algorithme AdaBoost, une première distribution de poids est construite sur l'ensemble d'entrainement. Le mécanisme de boosting sélectionne ensuite le classificateur de base qui donne le minimum d'erreurs. Ainsi, l'algorithme favorise le choix d'un autre classificateur qui est plus performant sur les données restantes au cours de la prochaine itération. Dans le domaine de la détection de personnes, AdaBoost a été utilisé par Viola et al.[VJS03] pour détecter les piétons dans une scène de rue.

## 2.5 Analyse du comportement humain dans des scènes individuelles

Il existe plusieurs études sur l'analyse du comportement humain dans des scènes individuelles. Les récentes études de Forsyth et al. [FAI+05], ainsi que de Pope [Pop10] s'intéressent à la reconnaissance des actions à partir de flux vidéos. D'autres travaux sur ce sujet, apparaissent dans [AC97, Bob97, Gav99, KKUG07, MHK06, TCSU08, WHT03]. Bobick [Bob97] a recours à une taxonomie pour la détection des mouvements, des actions et des activités. Il est à noter que nous définissons les actions et les activités de manière différente ; une action est un

ensemble de mouvements simples réalisés durant un laps de temps court alors qu'une activité est un ensemble d'actions qui se déroulent dans un laps de temps plus long (ex. se battre, garer sa voiture, etc.).

L'analyse du comportement dans une scène individuelle ne se limite pas à la reconnaissance des actions. Par exemple, Aggarwal et Cai [AC97], ainsi que Wang et al. [WHT03] ont publié une étude sur l'analyse de la structure du corps, sa détection ainsi que son suivi. Également, Moeslund et al. [MHK06] ont effectué des travaux sur la détection de la pose qui consiste à trouver la posture d'une personne dans une image. La détection d'individus ou de piétons [EG09, GT07, GLSG10], dont l'objectif est de localiser des personnes dans une image, est également liée à ce domaine de recherche. Nous citons finalement les travaux de [KKUG07] sur la détection des intentions et l'apprentissage par imitation.

Nous focalisons notre étude sur la reconnaissance d'actions humaines depuis la vidéo. Nous abordons les caractéristiques ou descripteurs utilisés pour la représentation des images ainsi que les méthodes de classification des actions. Nous étudions dans ce qui suit les méthodes de représentation des actions.

Les méthodes de reconnaissance d'actions se déroulent généralement en trois étapes : (i) Estimation du mouvement issu de l'action, (ii) Modélisation ou représentation de l'action, (iii) Reconnaissance de l'action. La modélisation de l'action est très importante car elle a pour but de représenter efficacement les actions. Elle doit être suffisamment riche pour permettre la détection de divers types d'actions.

L'aspect temporel est important lors de la réalisation d'une action. Ainsi, certaines représentations prennent clairement en compte la dimension temporelle, tandis que d'autres extraient les caractéristiques de chaque image individuellement dans la séquence. Dans ce dernier cas, les variations temporelles doivent être traitées lors de la classification.

Les méthodes de modélisation des actions peuvent être divisées en deux catégories : **globales** et **locales**.

(a) Les représentations globales portent sur les observations visuelles dans leur ensemble et sont obtenues de manière descendante : une personne est d'abord localisée dans l'image grâce à l'extraction de l'arrière-plan ou au suivi. Puis la région d'intérêt est encodée dans son intégralité, ce qui aboutit à un descripteur d'image. Ces représentations sont puissantes car elles permettent d'encoder la plupart des informations. Cependant, elles dépendent de l'exactitude de la localisation, de l'extraction de l'arrière-plan ou du suivi. De plus, elles sont plus sensibles à l'angle de la caméra, au bruit et aux occlusions. Lorsque le contexte applicatif permet de bien contrôler tous ces facteurs, la représentation générale donne des résultats satisfaisants.

(b) Les représentations locales décrivent les observations comme des ensembles de patchs (ou imagettes) indépendants. Le calcul des représentations locales est effectué de manière ascendante : les points d'intérêt spatiotemporels sont d'abord détectés, puis les patchs locaux sont calculés autour de ces points. Enfin, les patchs sont combinés en une représentation finale. Suite au succès initial des approches liées aux sacs de caractéristiques (bag of features), les chercheurs s'intéressent actuellement davantage aux corrélations entre les patchs. Les représentations locales sont moins sensibles au bruit et aux occlusions partielles, et elles ne nécessitent pas nécessairement un suivi ou une soustraction de l'arrière-plan.

Dans ce qui suit, nous discutons chacune de ces représentations globales.

### 2.5.1 Représentations globales

Les représentations globales permettent d'encoder une région d'intérêt (RdI) d'une personne au sein d'une image. Les RdI sont généralement obtenues en faisant une extraction de l'arrière-plan ou en effectuant un suivi. Les représentations globales les plus courantes utilisent les silhouettes, les bords ou le flux optique. Elles sont sensibles au bruit, aux occlusions partielles et aux variations de l'angle de vue. Pour résoudre en partie ces problèmes, des approches à base de *grille* consistent à diviser les observations en cellules, chacune d'entre elles encodant localement une partie des observations.

Ces approches ne tirent pas pleinement partie de la dimension temporelle des vidéos contrairement aux approches utilisant un volume spatiotemporel en 3 dimensions obtenu en empilant de nombreuses images à travers le temps (le temps étant la 3ème dimension). Ce type de représentation est réalisé en divisant une image à l'aide d'une grille, pouvant ainsi partiellement éviter les variations dues au bruit, les occlusions partielles et les changements d'angle. Dans la cas d'une vidéo, la grille est appelée grille spatiotemporelle. Chacune de ses cellules décrit l'observation de l'image au niveau local pendant un laps de temps.

Thurau et Hlavác [TH08] utilisent des histogrammes de gradients orientés (HOG, [SSJNFF08]) et s'intéressent aux bords de l'avant-plan en appliquant une factorisation matricielle non-négative. Lu et Little [LL06] appliquent une Analyse en composantes principales après avoir calculé le descripteur HOG, ce qui permet de réduire considérablement la dimensionnalité.

Les images d'historique et d'énergie du mouvement (Section 2.3.3) ainsi que les volumes spatiotemporels (Section 2.3.2) font également partie des représentations globales.

Les représentations globales permettent d'obtenir de bons résultats. Cependant, elles ne gèrent pas les occlusions et sont applicables uniquement dans les zones où le mouvement est déterminé de façon fiable. Les représentations locales proposent de venir à bout de ces inconvénients.

### 2.5.2 Représentations locales

Les approches appartenant à cette catégorie décrivent les observations comme une série de descripteurs locaux ou patchs. Il n'est pas nécessaire de procéder à une localisation exacte ni à une soustraction de l'arrière-plan. Les représentations locales sont généralement robustes, voire invariantes aux changements d'angle, à l'apparence des personnes et aux occlusions partielles.

Les patchs sont échantillonnés de manière dense (réparties de façon uniforme) ou bien au niveau de certains endroits spécifiques. Ces derniers sont les emplacements correspondant aux zones de mouvement. Les descripteurs locaux décrivent des petites fenêtres (2D) dans une

image, ou des cuboïdes (3D) dans un volume vidéo. Comme pour les représentations globales, les observations peuvent être regroupées localement dans une grille. En exploitant les corrélations spatiotemporelles entre les patchs, les actions peuvent alors être modélisées de façon plus efficace car seuls les patchs significatifs sont conservés.

Dans ce qui suit, nous décrivons deux types de représentation locale, qui sont les détecteurs de points d'intérêt spatiotemporels et les descripteurs locaux.

#### Détecteurs de points d'intérêt spatiotemporels

Les points d'intérêt spatiotemporels sont des emplacements dans l'espace et le temps qui reflètent des changements de mouvements soudains dans une vidéo.

Laptev et Lindeberg [LL03] ont étendu le détecteur de coins de Harris [HS88] en lui rajoutant la dimension temporelle. Les points d'intérêt spatiotemporels sont les points dont le voisinage local est soumis à une variation significative dans les plan spatial et le plan temporel. L'échelle du voisinage est automatiquement sélectionnée pour l'espace et le temps individuellement. Ces travaux ont été améliorés par Laptev et al. [LCSL07] pour compenser les mouvements relatifs de la caméra.

Ces méthodes comportent cependant un inconvénient : le nombre des points d'intérêt stables est relativement faible. Ce problème est traité par Dollár et al. [DRC+05], qui appliquent un filtre de Gabor aux dimensions spatiales et temporelles de manière individuelle. On ajuste le nombre de points d'intérêt en changeant la dimension spatiale et temporelle du voisinage dans lequel les minimas locaux sont sélectionnés.

#### Descripteurs locaux

Les descripteurs locaux caractérisent un patch d'une vidéo ou d'une image comme une représentation idéalement invariante au bruit, à l'apparence de la personne et aux occlusions de l'arrière-plan et éventuellement à la rotation et à l'échelle. La dimension spatiale et temporelle d'un patch est généralement déterminée par l'échelle du point d'intérêt. La Figure 2.11

montre des cuboïdes où l'on voit que les personnes effectuant des actions similaires génèrent des cuboïdes similaires. Liu et al. [LYSS12] combinent les cuboïdes ainsi que la similarité de l'apparence afin de constituer des descripteurs de haut niveau décrivant les actions.

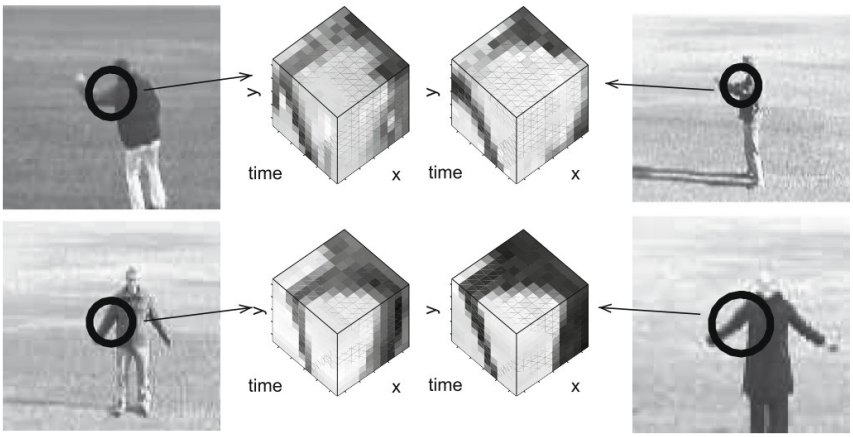



FIGURE 2.11 – Représentation des points d'intérêt spatiotemporels de Laptev et Lindeberg [LCSL07] pour des actions similaires exécutées par des personnes différentes

Les patchs peuvent également être décrits par des descripteurs locaux sous forme de grille. Ils intègrent les observations locales dans les cellules de la grille, ignorant ainsi les petites variations spatiotemporelles. Le descripteur SURF [BETVG08] est adapté à la 3D par Willems et al. [WTG08] sous le nom de eSURF (SURF étentdu). Laptev et al. [LMSR08] utilisent les descripteurs HOG et HOF (Histogramme de Flux Orienté). L'extension du HOG à la 3D est présentée par Kläser et al. [KMS08]. Les gradients 3D sont emboîtés en polyèdres réguliers. Il s'agit d'une extension de l'idée d'image intégrale en 3D qui permet de faire un échantillonnage dense et rapide du cuboïde à plusieurs échelles et localisations spatiotemporelles. Dans les travaux connexes de Scovanner et al. [SAS07], le descripteur SIFT [Low04] est lui aussi étendu à la dimension temporelle.

Messing et al. [MPK09] construisent des vocabulaires sur la trajectoire des points en mouvement. Chaque mot du vocabulaire peut être considéré comme un historique des magnitudes des vecteurs de mouvement. Des exemples de trajectoires sont illustrés dans la Figure 2.12.

Il n'est pas évident de comparer directement deux ensembles de descripteurs locaux car leur nombre peut être différent et ils sont généralement de dimensionnalité élevée. Par conséquent,

FIGURE 2.12 – Illustration des trajectiores extraites par l'approche de Messing et al. [MPK09]

un *livre de codes (codebook)* est souvent généré en regroupement des patchs et en sélectionnant soit le centre des groupes ou les patchs les plus proches du centre comme étant un *codeword*. Un descripteur local se définit comme un codeword qui contribue au codebook. Une image ou une séquence peut être représentée comme un sac de mots, c'est-à-dire un histogramme des fréquences des codewords.

### 2.5.3 Synthèse

Nous résumons les avantages et les inconvénients des approches basées sur les représentations globales et locales pour la reconnaissance d'actions humaines dans le Tableau 2.3.

| Méthode | Avantages | Inconvénients |
|---|---|---|
| Représentations globales | Bons résultats et coût d'exécution faible | Applicable uniquement dans le cas où les ROI sont déterminés de façon fiable. Ne traitent pas les occlusions |
| Représentations locales | Traitement des corrélations temporelles | Gestion des occlusions importantes ignorée |

TABLE 2.3 – Tableau synthétisant les avantages et inconvénients des méthodes de reconnaissance des actions humaines

Les représentations globales de l'image offrent généralement de bons résultats et celles-ci peuvent généralement être extraites facilement. Cependant, leur applicabilité est limitée aux scénarios dans lesquels les régions d'intérêt peuvent être déterminées de manière fiable. De plus, elles ne gèrent pas les occlusions.

Les représentations locales gèrent les occlusions de façon plus efficace. Des travaux préliminaires utilisent des sacs de représentation des caractéristiques, mais d'autres plus récents prennent en compte les corrélations spatiotemporelles entre les patchs. Néanmoins, la gestion d'occlusions plus importantes a été largement ignorée.

Après avoir exposé un état de l'art sur l'analyse du comportement humain dans une scène individuelle, focalisé sur la reconnaissance d'actions, nous présentons dans la section suivante un état de l'art axé sur l'analyse du comportement humain dans une scène de foule.

## 2.6 Analyse du comportement humain dans des scènes de foule

Dans les scènes de foule, trois types de problèmes sont posés couramment : (i) l'extraction de motifs de mouvement, (ii) la détection d'évènements et (iii) l'estimation des flux. Ces problèmes ne sont pas nouveaux et ont été abordés dans plusieurs études [CLK00, HTWM04, ZMR+08, BBE+08, MT08]. A travers cet état de l'art, nous décrivons les descripteurs ou types d'information exploités pour traiter chacun des trois problèmes afin d'aboutir à un ensemble d'informations caractérisant de façon commune ces 3 problèmes.

### 2.6.1 Extraction des motifs de mouvement

L'extraction des motifs de mouvement consiste à déterminer les habitudes de mouvement les plus fréquentes des objets [4] présents dans la vidéo. Ceci est fait en estimant en premier lieu

4. personnes, véhicules, etc

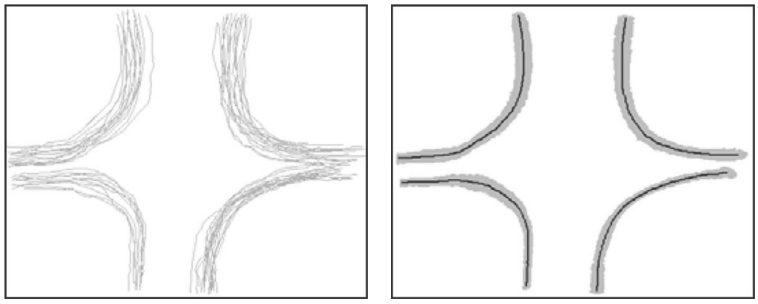
le mouvement ou les trajectoires des objets dans la scène. Ensuite, on applique des algorithmes capables d'en extraire les motifs de mouvements.

Nous distinguons deux types de scènes de foule dans cette problématique ; les scènes structurées et non structurées. Une scène est dite *structurée* quand les objets ne se déplacent pas de façon aléatoire et suivent une seule direction (ou modalité) de mouvement dans chaque région de la scène. Par exemple, une scène de course de 100 mètres est structurée car les coureurs ont la même direction de mouvement orientée vers la ligne d'arrivée. A l'opposé, une scène *non structurée* peut contenir des types de mouvement différents dans différentes régions de la scène. Par exemple, les passages piétons sont empruntés par les piétons dans deux sens différents, mais également franchis par des voitures dans différentes directions.

Nous classons les méthodes d'extraction des motifs de mouvement de la manière suivante : (i) les méthodes basées sur l'optimisation itérative, (ii) les méthodes basées sur l'adaptation en ligne, (iii) les méthodes hiérarchiques et (iv) les méthodes spatiotemporelles. Nous les développons dans les paragraphes suivants.

**Méthodes basée sur l'optimisation itérative**

Ces approches regroupent les trajectoires des objets en mouvement en utilisant des classificateurs simples tels que les K-moyennes. M. Hu et al. [HXF+06] génèrent des trajectoires en utilisant des algorithmes K-moyennes floues à partir de trajectoires des objets de l'avant-plan. Les trajectoires sont ensuite regroupées de façon hiérarchique et chaque modèle de mouvement est représenté par une chaine de Gaussiennes. La Figure 2.13 montre quatre motifs de mouvement extraits à partir d'un ensemble de trajectoires. Ces approches ont l'avantage d'être simples et efficaces. Toutefois, le nombre de motifs à extraire doit être spécifié manuellement et les trajectoires doivent être de longueurs (temporelles) égales et fixes, ce qui limite l'aspect dynamique. En plus, ces méthodes reposent sur l'hypothèse que les trajectoires individuelles sont correctes, ce qui n'est pas toujours le cas dans les scènes à haute densité de foule.



(a) (b)

FIGURE 2.13 – Quatre motifs en (b) de mouvements extraits depuis les trajectoires en (a) en utilisant l'approche de M. Hu et al. [HXF+06]

**Méthodes basées sur l'adaptation en ligne**

Ces approches intègrent de nouvelles trajectoires à la volée contrairement aux approches d'*optimisation itérative*. Cela est possible en utilisant un paramètre supplémentaire qui contrôle le taux de mises à jour.

Wang et al. [WTG06] proposent une mesure de similarité entre deux trajectoires qui permet de les regrouper, puis d'apprendre un modèle de scène à partir des groupes de trajectoires. La Figure 2.14 illustre l'étape de suivi des objets pour la construction des trajectoires ainsi que les motifs de mouvement estimés. Basharat et al. [BGS08] extraient des motifs de mouvement ainsi que des modèles de dimension et de mouvement pour chaque objet détecté. Ceci est réalisé en modélisant la loi de probabilité de la vitesse, la taille et la position sur chaque pixel de l'objet. Les modèles appris sont ensuite utilisés pour détecter des trajectoires ou objets ayant une trajectoire anormale.

Ces approches sont adaptées aux applications temps réel et aux scènes dont le nombre d'objets est variable dans le temps car ce dernier est mis à jour au fil du temps. Il n'est pas non plus nécessaire de maintenir une base de données d'apprentissage. Toutefois, il est difficile de choisir un critère pour l'initialisation des nouveaux groupes qui assure l'optimalité des résultats et empêche les valeurs aberrantes.

(a) (b)

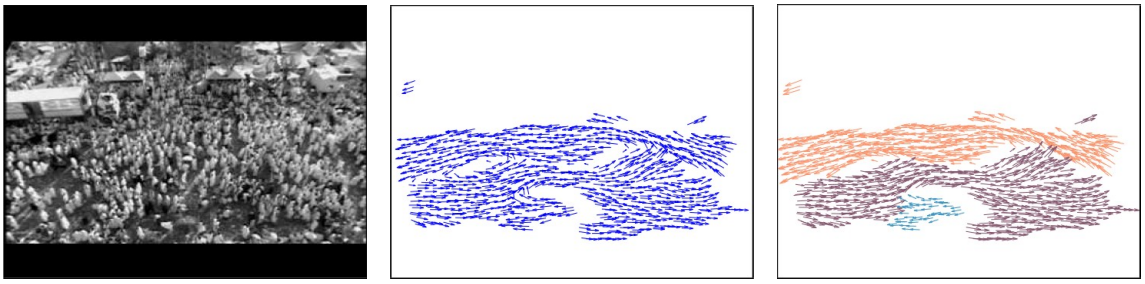
FIGURE 2.14 – Illustration de l'approche de Wang et al. [WTG06], (a) détection et suivi d'objets pour estimer les trajectoires, (b) motifs de mouvement estimés

**Méthodes hiérarchiques**

Ces approches considèrent une séquence vidéo comme la racine d'un arbre où les feuilles correspondent à des trajectoires individuelles.

W. Hu et al. [HAS08b] détectent les motifs de mouvement d'une séquence en regroupant son champ de vecteurs de mouvement. Chaque modèle de mouvement est constitué d'un groupe de vecteurs de flux optique qui est généré suite à des mouvements similaires (les auteurs parlent de processus de mouvement). La Figure 2.15 illustre le champ de vecteurs de mouvement et les motifs de mouvement détectés dans une scène de pèlerinage. Cependant, l'algorithme proposé est conçu uniquement pour les scènes structurées. Il nécessite aussi de spécifier le nombre maximum de motifs, qui doit être légèrement plus élevé que le nombre de motifs désirés ou réellement présents dans la scène. Zhang et al. [ZLL09] modélisent les trajectoires des véhicules et des piétons à l'aide d'un graphe, et y appliquent l'algorithme *graph-cut* afin de regrouper les motifs de mouvement.

Ces approches sont bien adaptées pour les techniques de la théorie des graphes (comme le flot maximal et de la coupe minimale). En outre, le regroupement multi-échelle permet un choix plus fin du nombre de groupes. L'inconvénient des méthodes hiérarchiques réside dans la qualité des groupes, qui dépend de la décision de fusion ou de division d'un ensemble et qui n'est pas ajustée le long de l'arbre.



(a) (b) (c)

FIGURE 2.15 – Approche de W. Hu et al. [HAS08b] appliquée à une vidéo de pèlerinage en (a), (b) champ de vecteurs de mouvement, (c) motifs de mouvement

**Méthodes spatiotemporelles**

Ces approches utilisent le temps comme une troisième dimension, et considèrent la vidéo comme un volume 3D contenant des points de coordonnées (x, y, t) (comme avec les descripteurs spatiotemporels vus dans la section 2.5).

Yu et Medioni [YM09] extraient les motifs de mouvement des véhicules dans des séquences vidéo filmées depuis les airs. Ceci est réalisé en utilisant une représentation en 4 dimensions $(x, y, v_x, v_y)$ du mouvement, avant d'appliquer le vote de tenseurs et la segmentation du mouvement. Lin et al. [LGF09] transforment la séquence vidéo dans un espace vectoriel en utilisant une représentation d'algèbre de Lie. Les modèles de mouvement sont ensuite appris en utilisant un modèle statistique appliqué à l'espace vectoriel. Gryn et al. [GWT09] introduisent la *carte de direction* comme une représentation qui capture la distribution spatiotemporelle de la direction du mouvement dans la séquence vidéo. Toutefois, la carte de direction est capable de capturer une seule orientation dominante ou modalité de mouvement pour chaque région de la scène. La Figure 2.16 illustre une carte de direction qui représente une voiture tournant à gauche.

**Méthodes de cooccurrence**

Ces méthodes tirent parti des résultats obtenus dans les domaines de la recherche de documents et du traitement du langage naturel. La vidéo est considérée comme un document et un motif de mouvement comme un sac de mots.

FIGURE 2.16 – Carte de direction correspondant à une voiture tournant à gauche

Rodriguez et al. [RAK09] proposent de modéliser le comportement des foules en utilisant un modèle de corrélation des sujets (CTM, ou Correlated Topic Model). Les motifs appris sont utilisés comme une connaissance a priori afin d'améliorer les résultats du suivi. Ce modèle utilise l'orientation du vecteur de mouvement, par la suite discrétisée en quatre directions de mouvement, comme une caractéristique de bas niveau. Ce travail est basé sur la division manuelle de la vidéo en sous-séquences courtes. Une étude plus approfondie est nécessaire afin de déterminer la durée pertinente des ces sous-séquences. Stauffer et Grimson [SG00] utilisent un algorithme de suivi en temps réel afin d'apprendre les motifs de mouvement des pistes obtenues. Ils appliquent ensuite un classificateur illustré dans la Figure 2.17 afin de détecter les évènements inhabituels. Ceux-ci sont définis par des mouvements qui ne respectent pas les motifs de mouvement préalablement extraits.

Grâce à l'utilisation d'une matrice de cooccurrence à partir d'un vocabulaire fini, ces approches sont indépendantes de la longueur de la trajectoire. Toutefois, la taille du vocabulaire est limitée pour conserver un regroupement efficace et la dimension temporelle est parfois négligée.

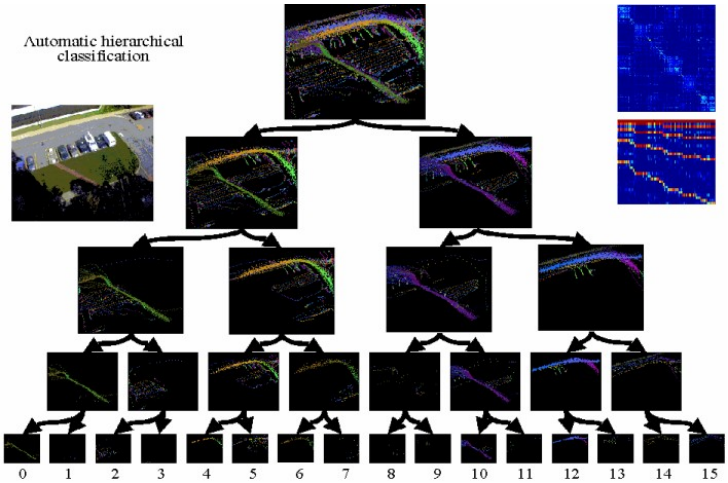



FIGURE 2.17 – Illustration du classificateur de Stauffer et Grimson [SG00]

**Synthèse**

Nous résumons les avantages et les inconvénients des approches d'extraction des motifs de mouvement dans le Tableau 2.4. L'analyse des motifs des mouvements en exploitant les trajectoires individuelles n'est pas adaptée pour les scènes de foule car la détection et le suivi des personnes y sont difficiles. Le flux optique est la méthode la plus adaptée et la plus utilisée car il permet d'estimer le mouvement d'une personne ou d'une foule indépendamment du nombre de personnes dans la scène. Nous remarquons également que la plupart des approches de l'état de l'art ne traitent que les scènes structurées où le déplacement dans la scène est organisé et uniforme. Cependant, on rencontre souvent des scènes non structurées dans des situations réelles. Ces scènes sont plus complexes car on peut y trouver plusieurs motifs de mouvement dans une même région.

Nous avons vu que certaines approches d'extraction des motifs de mouvement permettent de détecter des évènements inhabituels. Cependant, la problématique de détection d'évènements de foule est plus large. La section suivante présente un état de l'art de cette problématique.

| Méthode | Avantages | Inconvénients |
|---|---|---|
| Optimisation itérative | Efficace malgré sa simplicité. | Le nombre de groupes doit être spécifié manuellement. Trajectoires de longueurs égales. |
| Adaptation en ligne | Adaptée aux applications temps réel et aux scènes dont le nombre d'objets est variable dans le temps. Ne nécessite pas de maintenir une base de données d'apprentissage. | Le choix du critère d'initialisation optimal est difficile. |
| Méthodes hiérarchiques | Adaptées pour les techniques de la théorie des graphes. | La qualité des motifs dépend de la décision de fusion ou de division d'un ensemble et qui n'est pas ajustée ou réévaluée le long de l'arbre. |
| Méthodes de cooccurrence | Indépendantes de la longueur de la trajectoire. | Taille du vocabulaire limitée et la dimension temporelle est parfois négligée. |

TABLE 2.4 – Tableau synthétisant les avantages et inconvénients des approches d'extraction des motifs de mouvement

### 2.6.2 Détection d'évènements de foule

La détection d'évènements dans des vidéos de foule a attiré l'attention de nombreux chercheurs ces dernières décennies. Des études sur la détection d'évènements de foule [HTWM04, ZMR+08, BBE+08] sont à la disposition de la communauté scientifique. En règle générale, un système de détection d'évènements passe par les étapes suivantes [HAS08a] : (i) la détection de chaque objet en mouvement, (ii) le suivi des objets détectés et (iii) l'analyse de leurs vitesse et trajectoire pour détecter des évènements ou des activités.

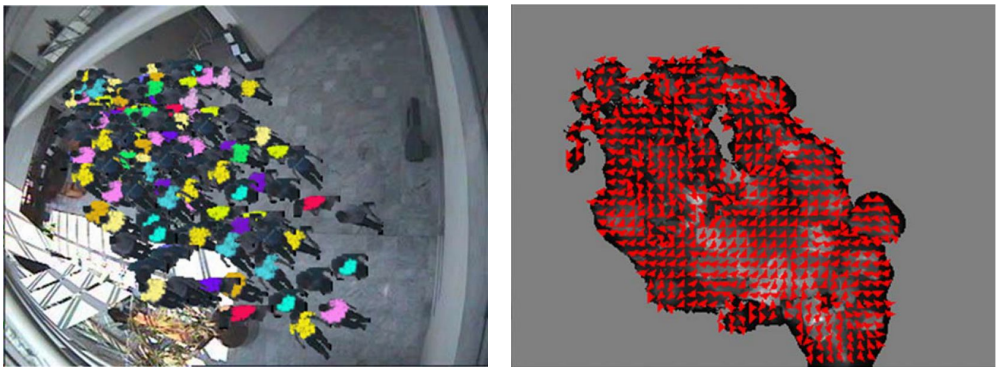
Cependant, il est très difficile de détecter chaque objet individuellement dans des scénarios complexes impliquant des scènes de foule dense. Les approches globales s'intéressant à l'ensemble de la scène et non à des individus distincts sont plus pertinentes. Les approches peuvent cibler deux objectifs : d'une part, elles estiment la densité de la foule, et d'autre part, elles détectent des évènements dans des environnements fréquentés.

Le premier objectif peut être abordé avec des méthodes reposant sur l'extraction des textures et sur l'analyse du coefficient de surface en mouvement [LCC01, MLHT04]. Ces méthodes

permettent d'obtenir une analyse efficace lorsque la foule est statique, mais ne peuvent pas détecter les évènements nécessitant une analyse dynamique de la scène. Certaines approches ont recours au flux optique [BV99, DYV95] pour assurer le suivi d'une foule ou d'individus stationnaires grâce à plusieurs caméras [CABT04].

Concernant le second objectif, où il est question de détecter des évènements se produisant au sein d'une foule, la méthodologie générale consiste à générer des modèles sur le comportement de la foule pour ensuite les utiliser pour inférer les évènements qui se produisent dans le flux vidéo. On y trouve deux types d'évènements : les évènements anormaux et les évènements sémantiques.

La plupart des approches se focalisent sur la détection d'évènements anormaux. Ces derniers sont considérés comme des cas aberrants et sont caractérisées par leur déviation des comportements typiques.



(a) (b)

FIGURE 2.18 – Illustration de l'approche d'Andrade et al. [ABF06]. (a) Foule simulée, (b) Données de flux optique simulées

### Méthodes de détection d'évènements anormaux

Il est possible d'utiliser une approche de détection des motifs de mouvement (voir Section 2.6.1) afin de modéliser le mouvement de la foule dans des situations normales. Un évènement anormal (ou inhabituel) est alors défini comme étant un mouvement qui ne respecte pas les motifs de mouvement extraits dans des situations normales. Cette approche permet de détecter

des évènements impliquant des déplacements dans des directions inhabituelles (ex. une voiture qui roule à contre sens). Cependant, le type d'évènement détecté est très limité car il s'agit des situations qui ne respectent pas les motifs de mouvement, qui sont des contre sens généralement. Ceci peut aussi amener des confusions car une voiture qui effectue un créneau n'est pas un évènement inhabituel.

Plusieurs approches proposent de détecter des comportements anormaux sans utiliser les motifs de mouvement. Andrade et al. [ABF06] combinent les modèles de Markov cachés avec l'analyse en composantes principales des vecteurs du flux optique pour détecter des scénarios d'urgences. Cependant, les expérimentations ont été portées sur des données simulées tel qu'illustré dans la Figure 2.18. Les humains sont simulés par des pantins noirs avec une chemise colorée. La simulation consiste à déplacer l'image de chaque pantin dans la scène. Cependant, la simulation ne permet pas de gérer les collisions entre les pantins et la lumière n'a pas d'effet sur eux (pas d'ombre et pas d'illumination du pantin).

Wu et al. [WMS10] utilisent la dynamique des particules Lagrangiennes pour détecter les instabilités du flux. Cette méthode est efficace pour la segmentation des grandes densités de foules (marathons, évènements politiques et religieux, etc.). Ihaddadene et Djeraba [ID08] détectent les situations d'écroulement en se basant sur une mesure qui décrit le degré d'organisation ou de désordre des vecteurs de flux optique. Cette approche fonctionne sur des zones unidirectionnelles (ex. escaliers automatiques). Ramin et al. [MOS09] utilisent le flux optique pour détecter les comportements anormaux dans la foule en utilisant un modèle de force sociale. Kratz et Nishino [KN09] déterminent les comportements-types se produisant dans des scènes de foule en modélisant la variation du mouvement avec des volumes spatiotemporels locaux. Ce modèle statistique est ensuite utilisé pour détecter des comportements anormaux.

L'approche proposée par Adam et al.[ARSR08] permet de détecter des évènements inhabituels en analysant des régions spécifiques dans une séquence vidéo par le biais de "moniteurs" illustrés dans la Figure 2.19(a). Chaque moniteur extrait des observations locales de bas niveau associées à la région correspondante comme la magnitude moyenne du flux optique (voir la Figure 2.19(b)). Un moniteur utilise un tampon cylindrique pour calculer la probabilité d'une

observation en cours par rapport à des observations précédentes. Les résultats donnés par plusieurs moniteurs sont ensuite intégrés pour alerter l'utilisateur d'un comportement anormal.

(a) (b)

FIGURE 2.19 – Illustration de l'approche d'Adam et al. [SG00]. (a) Une scène typique contenant des moniteurs représentés par des points rouges, (b) la magnitude moyenne du flux optique observée par chaque moniteur

L'inconvénient majeur de ces approches est qu'elles ne fournissent pas de sémantique car elles ne permettent pas de catégoriser les évènements. Le deuxième type d'approche se focalise sur la détection d'évènements sémantique.

### Méthodes de détection d'évènements sémantiques

Les approches de détection d'évènements sémantiques proposent de détecter une plus large palette d'évènements et de leur donner un nom plus expressif tel que course, combat, chute, etc.

Dans cette catégorie, Utasi et al. [UKS09] proposent une approche reposant sur des descripteurs statistiques afin de détecter trois types d'évènements : activité régulière, course et séparation. L'approche commence par extraire l'arrière-plan pour ensuite calculer le flux optique sur les pixels de l'avant-plan. Ces vecteurs ainsi obtenus sont utilisés pour modéliser un mélange gaussien de 4 dimensions (x, y, vx, vy) (position et vitesse) dans les vidéos où la foule marche et reste constamment en groupe. La Figure 2.20 représente les moyennes des gaussiennes du mélange : les lignes rouges représentent la direction et la magnitude moyenne dans la position moyenne, tandis que les ellipses blanches sont proportionnelles aux variances des positions.

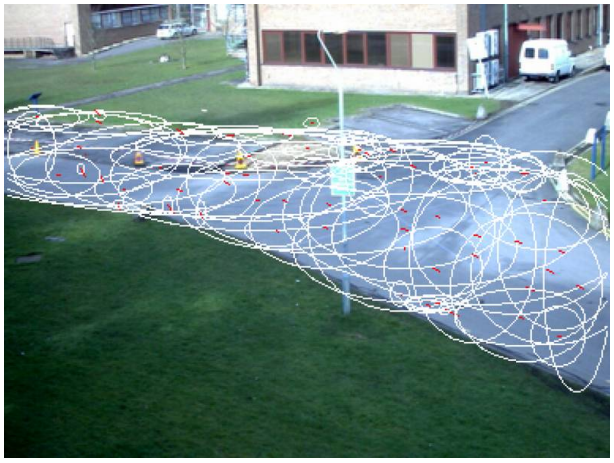


FIGURE 2.20 – Illustration des moyennes des gaussiennes d'un mélange Gaussien à quatre dimensions (x, y, vx, vy). Les lignes rouges représentent la direction et la magnitude moyenne dans la position moyenne, tandis que les ellipses blanches sont proportionnelles aux variances des positions

Ces modèles sont ensuite utilisés pour estimer une mesure de similitude $M_{sim}$ (similitude d'adhésion) et une mesure de dissemblance $M_{issim}$ (dissemblance d'adhésion) qui exprime la similarité entre un descripteur (flux optique ou magnitude de mouvement) par rapport aux modèles. Les auteurs expriment ensuite un évènement pour chaque sous-séquence à l'aide de ces deux valeurs. L'évènement régulier est exprimé en multipliant les deux mesures de similitude par rapport aux deux modèles. L'évènement course est exprimé en multipliant la mesure de dissimilitude du flux optique par la mesure de similitude de la magnitude. L'évènement séparation est exprimé en multipliant la mesure de dissimilitude du flux optique par la mesure de similitude de la magnitude. L'évènement donnant la plus grande valeur est celui qui est retourné pour la sous-séquence traitée.

Chan et al. [BMV09] proposent une approche reposant sur des propriétés globales afin de détecter 6 types d'évènements : marche, course, séparation, dispersion locale, évacuation et fusion. Elle consiste à modéliser le flux de la foule à l'aide du modèle de textures dynamiques. Une texture dynamique est un modèle qui traite une vidéo comme un échantillon d'un système linéaire dynamique. Bien que cette approche permette de détecter plusieurs évènements, elle est gourmande en calcul et ne permet pas le chevauchement des évènements (par exemple, les évènements marche et course peuvent se dérouler en même temps que l'évènement séparation).

**Synthèse**

Le Tableau 2.5 reprend les avantages et les inconvénients des méthodes de détection d'évènements anormaux et sémantiques. Les approches de détection d'évènements anormaux, bien que plus simples à mettre en œuvre, ne donnent pas de sémantique aux situations anormales. Les approches de détection d'évènements sémantiques permettent de détecter différents types d'évènements. Cependant, elles ne gèrent pas les situations où deux évènements se produisent au même instant dans la scène.

| Méthode | Avantages | Inconvénients |
|---|---|---|
| Détection d'évènements anormaux | Peuvent être dérivées des approches d'extraction de motifs de mouvement | Détectent des situations inhabituelles sans leur donner de sens |
| Descripteurs statistiques (évènements sémantiques) | Supporte des greffons pour détecter plus d'évènements | Détecte un faible nombre d'évènements. Pas de détails sur l'utilisation des greffons |
| Propriétés globales (évènements sémantiques) | Détecte 6 types d'évènements différents | Ne gère pas le chevauchement d'évènements |

TABLE 2.5 – Tableau synthétisant les avantages et les inconvénients des approches de détection d'évènements de foule

Nous avons expliqué au début de cette section que certaines approches d'analyse d'évènements de foule effectuent également une estimation de la densité de la foule (ce qui est le cas pour [UKS09]) qui nous permet d'estimer le nombre de personnes dans une image. Ceci fait partie d'une problématique plus large qui est l'estimation des flux. Nous présentons un état de l'art de cette problématique dans la section suivante.

### 2.6.3 Estimation des flux

De nombreuses approches d'estimation des flux par comptage des personnes ont été proposées dans la littérature. Le problème est souvent simplifié par l'utilisation d'une caméra zéni-

thale (verticale) [ALCC09], une caméra frontale [ZDC09] ou une configuration multi-caméras [YGBG03]. On peut diviser les approches en cinq catégories : (i) Méthodes basées sur l'analyse des trajectoires de mouvement, (ii) Méthodes basées sur l'analyse des contours, (iii) Méthodes basées sur les modèles, (iv) Méthodes basées sur la stéréovision et (v) Méthodes spatiotemporelles.

**Méthodes basées sur l'analyse des trajectoires de mouvement**

Ces méthodes sont appliquées en deux étapes. La première concerne la détection, dans une scène, des régions en mouvement correspondant principalement à des individus. La seconde a recours aux résultats relatifs à la détection permettant de reconstruire la trajectoire des objets en mouvement à travers le temps. L'analyse de la trajectoire permet d'identifier et de compter les personnes franchissant une ligne virtuelle ou une zone prédéfinie [LTR$^+$05], [ZC07], [XWLZ07].

Xu and al. [XWLZ07] proposent une méthode rapide de comptage utilisant un modèle de scène, un modèle humain et un modèle de terrain en prenant en compte les contraintes dans les modèles. Le modèle de scène est défini grâce à des coordonnées homogènes dont les paramètres peuvent être calculés en utilisant l'approche décrite dans [ZWW05] ou en utilisant les paramètres de la caméra fournis par son fabriquant.

Les modèles humains et de terrain sont des modèles de perspective basés sur le postulat que la caméra est fixée verticalement, tel qu'illustré dans la Figure 2.21. Cette dernière représente une tête humaine sous forme de sphère et un corps sous forme de cylindre. Les différents paramètres et coordonnées répondent à une géométrie triangulaire optimisée.

Les auteurs ont ensuite recours à la méthode de soustraction de blocs sur l'arrière-plan et à l'opération consistant à remplir les espaces vides. Les blobs qui correspondent aux personnes en mouvement sont extraits comme indiqué dans la Figure 2.22. Puis une segmentation grossière est effectuée afin de détecter les individus isolés en fonction du nombre de pixels dans un blob. Par la suite, une segmentation affinée est effectuée pour détecter les groupes de personnes grâce

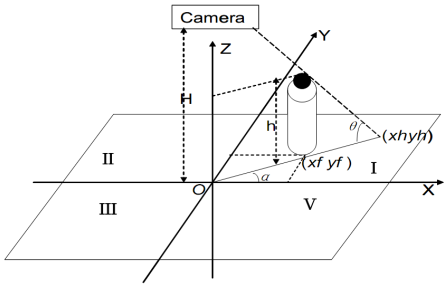
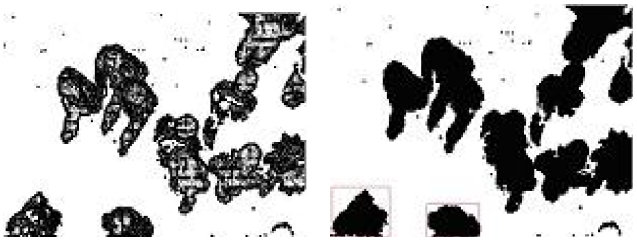



FIGURE 2.21 – Modèle humain avec caméra verticale

à la méthode de binarisation de Niblack. La Figure 2.23 montre les étapes de segmentation grossière et affinée. Les personnes détectées sont ensuite suivies.



(a) (b)

FIGURE 2.22 – Détection des blobs : (a) Soustraction de l'arrière-plan. (b) Remplissage des vides

(a) (b)

FIGURE 2.23 – Résultats de la segmentation dans une image : (a) Blobs détectés. (b) Résultats de la segmentation

Le système aboutit à de bons résultats et gère jusqu'à 15 piétons dans une même image en niveau de gris. Cependant, le calcul des paramètres doit être effectué manuellement pour chaque configuration de caméra et les hypothèses concernant la position de la caméra et la couleur de la tête sont restreintes.

Zhang et Chen [ZC07] comptent des personnes à partir de séquences vidéo capturées par une caméra monoculaire située à un angle inférieur à 45° en direction du sol. Cela entraîne une variation, dans le sens de la profondeur (ou axe Z), de la taille humaine dans une image qui permet de déduire la profondeur d'une personne selon sa hauteur.

Cette approche se divise en trois principaux modules : (i) la détection du mouvement, (ii) la segmentation humaine multiple et (iii) le suivi des groupes.

Le premier module, à savoir la détection du mouvement, estime les pixels du premier plan. Un modèle gaussien [WADP97a] est d'abord utilisé pour détecter les pixels en mouvement. Puis les ombres sont supprimées en appliquant l'algorithme décrit dans [MJD+00].

Le deuxième module est la segmentation humaine. Son rôle consiste à discriminer les personnes depuis le premier plan. L'hypothèse prise en compte est que, la plupart du temps, la tête est visible. Ainsi, les auteurs peuvent améliorer et optimiser la méthode de segmentation décrite dans [ZN04]. Elle consiste à détecter les têtes préalablement définies comme des points culminants locaux au sein du premier plan, comme illustré dans la Figure 2.24. Les fausses détections sont ainsi évitées grâce à l'utilisation d'une silhouette de projection verticale. Si la valeur de la projection correspondant au haut de la tête d'un candidat dépasse un certain seuil, elle est considérée comme la partie supérieure réelle de la tête. Le seuil est relatif à la taille humaine dans une image et n'est pas facile à déterminer. Dans la mesure où la caméra est placée selon un angle inférieur à 45° en direction du sol, la hauteur de l'image est plus élevée lorsque la personne est proche de la caméra. Ce problème peut être résolu en paramétrant la taille la plus haute et la plus basse dans une image. Pour ce faire, il faut utiliser l'interpolation linéaire qui permet de renvoyer la hauteur de l'image depuis n'importe quel emplacement de la scène.

Le dernier module est le suivi de groupe pour le comptage des individus. La segmentation humaine indique le nombre de personnes dans un groupe. Cependant, en cas d'occlusion partielle ou complète, on aboutit à un résultat incorrect lors de la segmentation. Par conséquent, le suivi de groupe permet d'enregistrer l'historique du nombre de personnes dans un groupe. Dans le cadre de cette approche, imaginons qu'il n'y ait pas d'occlusion complète, la segmentation humaine peut donc aboutir à un résultat correct sauf pour quelques images (par exemple,

(a) (b) (c)

FIGURE 2.24 – Segmentation humaine multiple. a) Image d'origine. b) Premier plan. c) Analyse de la bordure dans une région

3 images sur 10 ont été mal segmentées à cause d'une occlusion). Ainsi, les coefficients de confiance sont introduits afin d'améliorer le suivi. Le module de suivi est capable de fusionner, diviser et superposer des groupes.

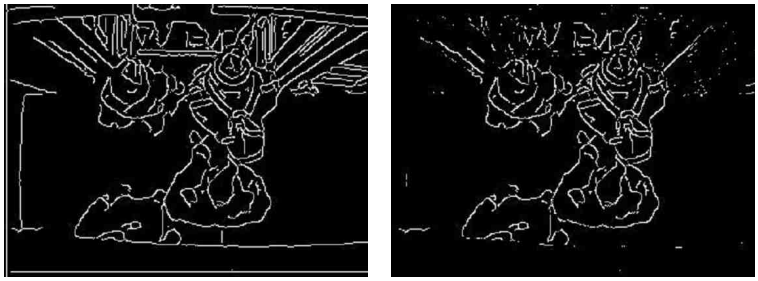
Cette approche a été validée dans la base de données PETS'2002 et a obtenu un taux d'erreur inférieur à 5%. Les expériences montrent que l'application de cette méthode est rapide. Néanmoins, cette approche implique que l'arrière-plan soit statique puisqu'une seule gaussienne est utilisée. Par ailleurs, elle est également sensible aux changements soudains de luminosité. Enfin, les hypothèses de la caméra sont très restrictives et l'interpolation entre la position de l'image et la profondeur réelle doit être définie pour chaque scène.

**Méthodes basées sur l'analyse des contours**

Ces méthodes consistent à extraire des objets d'intérêt dont les contours ont une forme et une organisation particulières. Ces objets sont ensuite comptés. Par exemple, une tête peut être considérée comme une région d'intérêt dont le contour est de forme circulaire [BCS07], [GBJ+07], [YCSX08], [LTJS12].

Bozzoli et al. [BCS07] proposent une approche qui compte les individus dans des environnements très fréquentés. Elle a recours à une caméra de faible coût installée verticalement au plafond. L'approche commence par estimer le mouvement par le biais des images de gradient, ce qui permet ensuite de déterminer l'avant-plan.

La détection du mouvement basée sur une image de gradients combine un modèle de l'arrière-plan tel que la moyenne mobile appliquée aux images de bord, et la soustraction de l'arrière-plan dans l'image de bord actuelle. La Figure 2.25 montre les bords statiques et mobiles d'une image tandis que la Figure 2.25(b) montre les bords de l'avant-plan et la carte des bords correspondante dans la Figure 2.25(a).



(a) (b)

FIGURE 2.25 – Bords statiques et mobiles d'une image. (a) Carte des bords statiques. (b) Bords mobiles

Après avoir obtenu les images de bord, les segments linéaires reliés sont regroupés et considérés comme faisant partie des mêmes objets. La Figure 2.26 montre les résultats de cette étape pour lesquels seuls les segments retenus sont visibles.

FIGURE 2.26 – Résultat du regroupement des segments reliés

Les vecteurs de flux optique sont ensuite calculés en utilisant l'algorithme vu en Section 2.2.3. Cependant, certaines erreurs apparaissent et sont corrigées lors du regroupement des segments reliés. Enfin, le comptage est effectué en calculant le nombre de pixels dans une région prédéfinie ayant la forme d'un polygone fermé et relié.

Bien que ce système soit efficace, il requiert l'initialisation manuelle de certains paramètres comme le $VPR$. La méthode d'extraction de l'arrière-plan nécessite davantage de tests et de validations.

**Méthodes basés sur des modèles**

Ces méthodes estiment les flux en comptant les régions dans les images traitées qui correspondent à des modèles prédéfinis [SLBS06, GBTT08]. Ces modèles représentent des personnes et sont soit des modèles de caractéristiques soit des modèles d'apparence.

Sidla et al. [SLBS06] proposent une approche capable de détecter, suivre et compter les individus dans des situations de foule. La Figure 2.27 montre les résultats obtenus par le module de suivi dans le cas d'un scénario dans le métro. La ligne bleue est appelée "ligne virtuelle" et est utilisée pour effectuer le comptage. Les personnes sont détectées en appliquant un filtre de région d'intérêt (ROI - Region of Interest) conçu à partir du modèle de l'arrière-plan (voir Section 2.2.2), et le détecteur de forme de type Ω. La trajectoire des individus est maintenue grâce à un filtre de Kalman. Le comptage est effectué en utilisant une entrée virtuelle prédéfinie et des heuristiques simples basées sur la trajectoire.

FIGURE 2.27 – Détection et suivi des personnes sur un quai de métro

La détection des personnes nécessite tout d'abord de détecter les contours. Il existe une méthode rapide pour déterminer les contours consistant à détecter les candidats correspondant à des formes de type Ω au sein des ROI selon les principes décrits par Zhao et Nevatia [ZN03]. Le contour d'une personne est représenté comme un modèle en 23 points. L'angle local d'un point du modèle se définit comme la direction du sommet vers la droite lorsque le contour est

traversé dans le sens des aiguilles d'une montre. Le processus de détection applique d'abord un détecteur de bord de Canny à l'image d'entrée masquée par une ROI. Une carte des angles $R$ est directement établie à partir de $O$ en calculant l'orientation du pixel local sur cette carte. Puis, pour chaque emplacement possible $(xs, ys)$ de la forme de référence sur l'image d'entrée, une fonction de coût $S$ est calculée. La Figure 2.28 montre les principales étapes de l'algorithme de détection.

Les résultats sont ensuite analysés en adaptant des modèles actifs d'apparence (ASM - Active Shape Models) afin de ne retenir que les candidats qui correspondent bien à des modèles prédéfinis.

FIGURE 2.28 – L'algorithme de correspondance de forme de Zhao. Les images s'affichent dans le sens des aiguilles d'une montre, de la droite vers la gauche : image d'entrée, calcul de la fonction de coût, tête finale des candidats, image de l'orientation des bords et carte étendue de l'image des gradients $O$

Dans le cadre du comptage, une heuristique est utilisée pour une entrée virtuelle $G$ en définissant les conditions suivantes pour une image vidéo $n$ :

1. Un point $p$ de trajectoire $T$ se trouve dans l'entrée virtuelle $G$.
2. Aucun autre point de $T$ ne s'est trouvé dans $G$ pour les images $i$, pour tout $i < n$.
3. Aucun autre point de la trajectoire ne s'est trouvé dans $G$ pour les $k$ dernières images.

Si les trois conditions sont maintenues, le comptage du piéton associé à $G$ est incrémenté. La troisième condition permet d'éviter qu'un individu ne soit compté plusieurs fois si celui-ci est associé à plusieurs trajectoires. $k$ dépend du nombre d'images par seconde dans la vidéo et

est généralement fixé à 5 pour les données de sortie. La direction d'une personne est basée sur la position du point de départ $T$ relatif à $G$.

Bien que cette approche soit très efficace, elle n'est pas rapide et requiert une saisie manuelle des paramètres.

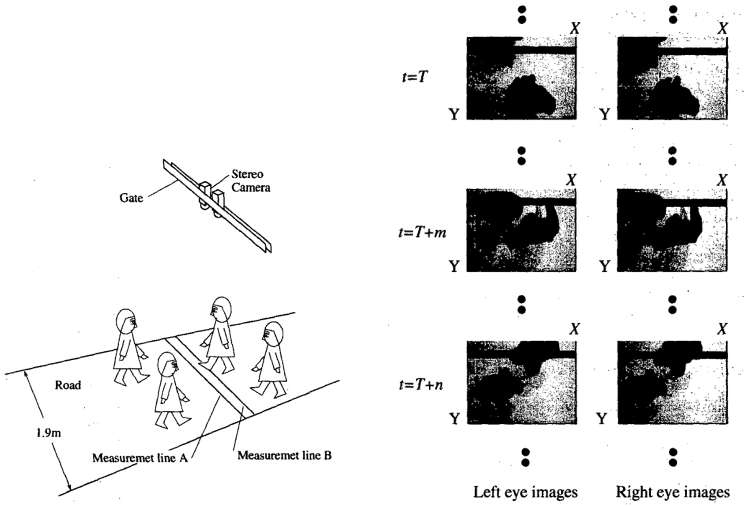
Bien que les approches basées sur les modèles permettent de détecter et de suivre des personnes en plus du comptage, elles nécessitent généralement une base d'apprentissage importante et/ou peuvent engendrer des problèmes liés à la généralisation des modèles.

#### Méthodes basées sur la stéréovision

Ces approches ont recours à des informations détaillées fournies par plusieurs caméras pour la détection et le comptage d'individus [Bey00, TYOY99, YK08, vOBK11].

L'approche de Terada and al. [TYOY99] consiste à compter les personnes passant une porte en utilisant une caméra stéréo. Celle-ci est accrochée au plafond au-dessus de la porte comme illustré dans la Figure 2.29(a), tandis que la Figure 2.29(b) montre quelques images obtenues. L'axe optique est fixé de façon à ce que l'observation des personnes se fasse uniquement au-dessus de leur tête. La disposition de ce système fait que les données relatives aux images des passants ne se chevauchent pas sur l'image obtenue en cas de foule. De plus, la hauteur et la largeur de chaque passant sont mesurées grâce à l'algorithme de triangulation.

La première étape pour le comptage des passants est d'obtenir une série d'images d'entrée correspondant à l'œil droit et à l'œil gauche de la caméra stéréoscopique. Puis, sur l'image correspondant à l'œil gauche, ne sont sélectionnés que les pixels se trouvant sur la ligne de comptage. Cette ligne est fixée sur le sol au niveau de l'angle droit selon la direction vers laquelle se déplacent les passants. Ces pixels sont ensuite disposés le long de l'axe temporel dans une image spatiotemporelle. De la même manière, depuis l'image correspondant à l' œil droit, les pixels se trouvant sur la ligne de calcul sont sélectionnés et transformés en une image spatiotemporelle. L'image spatiotemporelle générée par les images de l'œil droit et celle générée




(a) (b)

FIGURE 2.29 – Illustration d'un système de comptage stéréoscopique (a) Système de comptage où la caméra stéréo se trouve au-dessus de la porte, (b) Echantillons d'images obtenues avec une caméra stéréo

par les images de l'œil gauche sont mises en correspondance afin d'obtenir des données de forme tridimensionnelles.

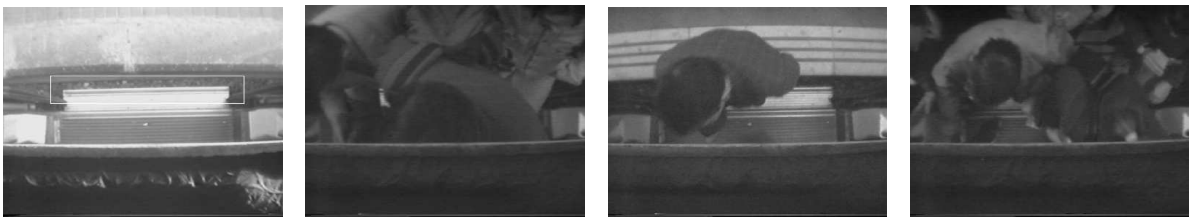
Les personnes entrantes et sortantes peuvent être comptabilisées grâce aux données de comptage projetées sur l'image spatiotemporelle comportant des informations sur chaque direction de déplacement et chaque hauteur. La direction est estimée à partir de la ligne de calcul. La position des points d'intérêt entre les images spatiotemporelles indique la direction de ce point. Le comptage est ainsi effectué grâce à l'analyse de la vitesse de passage des personnes. Si une personne marche lentement, la zone dans laquelle se trouvent les données est large autour de l'axe temporel. Si elle marche à vive allure, la zone de données est mince. Par conséquent, le comptage des passants peut être réalisé, permettant à la mise en correspondance de définir la vitesse de déplacement.

Les méthodes qui utilisent une représentation tridimensionnelle de la scène permettent de détecter les objets cachés, les transformant ainsi en méthodes robustes capables de gérer des situations complexes. Néanmoins, l'utilisation de plusieurs caméras indépendantes n'est pas encouragée car l'étape de calibration et le temps de calcul sont très longs.

**Méthodes spatiotemporelles**

Ces méthodes se basent sur la construction d'une carte spatiotemporelle en empilant à travers le temps les pixels d'une ligne de comptage prédéfinie. Des modèles statistiques sont ensuite utilisés pour définir le nombre de personnes franchissant la ligne de comptage et pour analyser les incohérences entre les cartes spatiotemporelles afin de déterminer la direction [AMN01, BMB08, YR06, CGZT09].

Albiol et al. [AMN01] s'intéressent au comptage des personnes qui montent ou descendent d'un train sans restriction en termes de fréquentation ou de changements de luminosité grâce à un algorithme différé qui s'exécute à des instants spécifiques. Cette approche comporte trois étapes : la première étape a lieu lorsque les portes sont fermées et que la caméra est recouverte. Le système y est inactif et attend l'ouverture des portes.



(a) (b) (c) (d)

FIGURE 2.30 – Echantillons d'images lors de l'acquisition. (a) Aucune personne, (b) Personne isolée, (c) et (d) Situations de foule

Puis l'étape d'acquisition commence lorsque les portes s'ouvrent et que les gens montent ou descendent du train comme indiqué dans la Figure 2.30. Divers empilements de personnes traversant une ligne virtuelle sur la porte sont ainsi établis. Les empilements sont respectivement de couleur blanche, gradient et noire comme indiqué dans la figure 2.31.

La dernière étape est celle du comptage qui se déclenche lorsque les portes se ferment. Il s'agit de traiter les empilements afin de récupérer le nombre de passants. La première opération consiste à effectuer une détection de présence en appliquant l'extraction de l'arrière-plan. Puis les personnes sont isolées grâce à la segmentation de l'image d'avant-plan en prenant en compte divers scénarios de passage (passage d'une personne, deux personnes en même temps, etc.)

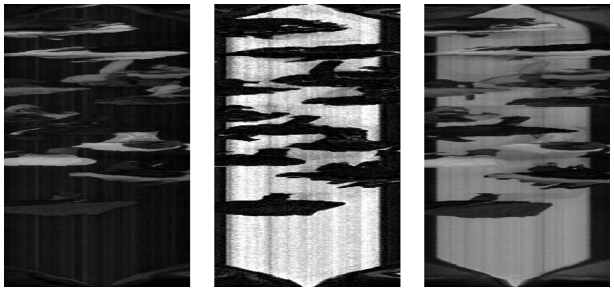


FIGURE 2.31 – Exemple d'empilement. De droite à gauche : empilements noirs, gradients, et blancs

La direction d'une personne est trouvée en estimant le mouvement de son barycentre grâce à une méthode de calcul du flux optique. La vitesse obtenue à partir de l'équation du flux optique donne une valeur par pixel empilé.

Les méthodes spatiotemporelles ont l'avantage d'être faciles et rapides à appliquer. Cependant, les travaux basés sur ces méthodes n'apportent pas des solutions concrètes permettant d'interpréter certains cas. Ainsi, lorsqu'une personne s'arrête sur la ligne virtuelle, un énorme blob se forme et est interprété, à tort, comme étant un groupe de personnes.

**Synthèse**

Nous synthétisons les avantages et les inconvénients des différents méthodes d'estimation des flux dans le Tableau 2.6. Les méthodes spatiotemporelles sont celles qui offrent le meilleur compromis résultats/vitesse d'exécution. Cependant, elles ont deux défauts significatifs. Le premier est qu'elles n'interprètent par correctement les personnes qui sont immobiles sur la ligne de comptage, les confondant avec des groupes de personnes. Le deuxième est qu'elles n'estiment pas le flux en temps réel mais nécessitent que la carte spatiotemporelle soit suffisamment grande.

Cette synthèse clôt l'état de l'art sur l'analyse du comportement humain dans des scènes de foule. La section suivante conclut ce chapitre en résumant les points clés abordés dans ce chapitre et en présentant nos contributions par rapport à l'état de l'art.

| Méthode | Avantages | Inconvénients |
|---|---|---|
| Analyse des trajectoires | Rapide et obtient de bons résultats | Non adaptées aux situations réelles (l'arrière-plan doit être statique). Nécessitent un paramétrage manuel. |
| Analyse du contour | Bon résultats | Initialisation manuelle de certains paramètres. |
| Modèles | Permet de détecter et de suivre les personnes en plus du comptage | Nécessitent une base d'apprentissage importante. Nécessitent une saisie manuelle des paramètres. |
| Stéréovision | Offrent la meilleure précision | Nécessitent une calibration. Gourmandes en calcul. |
| Spatiotemporelles | Rapides, simples et donnent de bons résultats | Détectent les personnes immobiles sur la ligne de comptage. N'estiment pas les flux en temps réel. |

TABLE 2.6 – Tableau synthétisant les avantages et inconvénients des approches d'estimation des flux

## 2.7 Conclusion

Ce chapitre a présenté un état de l'art portant sur l'analyse du comportement humain depuis la vidéo. Nous avons distingué deux types de scènes selon le nombre de personnes présentes.

Le premier type concerne les vidéos qui contiennent une personne qui est en train d'effectuer une action. Ce genre de scène est appelé scène individuelle. Plusieurs problèmes sont abordés dans le traitement de ce type de vidéos, parmi eux nous avons cité la reconnaissance des personnes, l'identification des personnes ou la reconnaissance des actions humaines. Nous avons donné un état de l'art concis sur le problème de la reconnaissance des actions humaines. Les approches traitant ce problème suivent généralement deux types d'approches distinctes : les approches globales et les approches locales qui étaient plus performantes. Nous proposons une nouvelle approche de description des actions basées sur le flux optique comme représentation locale. Notre approche analyse l'orientation et la vitesse du mouvement à travers le modèle directionnel et le modèle de magnitude respectivement. Nous montrons l'intérêt du modèle directionnel pour la reconnaissance d'actions et l'analyse des scènes de foule par la suite.

Le deuxième type de vidéos concerne celles qui capturent une scène contenant un nombre important de personnes. Ce genre de scène est aussi appelé scène de foule. Plusieurs problèmes sont abordés dans ce type de vidéos, parmi eux nous citons l'extraction des motifs de mouvement, la détection d'évènements et l'estimation des flux. Nous avons également présenté un état de l'art pour chaque problème. Nous remarquons que le flux optique est l'information la plus pertinente qui est utilisée dans la plupart des approches notamment quand le nombre de personnes est important car elle permet d'avoir de bons résultats dans un laps de temps réduit. Nous nous inspirons de ce résultat et nous estimons le mouvement comme une caractéristique de bas niveau dans notre approche.

La plupart des approches d'extraction des motifs de mouvement ne peuvent pas être appliquées dans des situations réelles à cause des contraintes qu'elles supposent. Certaines supposent que le nombre de personnes présentes dans une image ne soit pas très important et utilisent ainsi l'extraction de l'arrière-plan pour détecter les objets en mouvement. D'autres supposent que les modalités de mouvement soient uniques dans chaque région de la scène et ignorent la présence de mouvements dans des directions variées. Notre approche tente d'aborder le problème d'extraction des motifs de mouvements sans contraintes particulières, en utilisant le mouvement comme information de bas niveau et en estimant un modèle directionnel dans le niveau intermédiaire. Nous proposons au niveau sémantique un algorithme qui permet d'extraire les motifs de mouvement à partir du modèle estimé.

Quant à la détection des évènements de foule, nous avons noté l'existence de deux types d'approches ; celles qui détectent des évènements anormaux et celles qui détectent les évènements sémantiques. Le premier type d'approches offre une solution d'appoint dans les cas où la notion d'anormalité est clairement définie. Cependant, il est très compliqué de définir ce qui anormal à cause de l'imprévisibilité et à l'indéterminisme du comportement des personnes dans une scène de foule. Les approches de la deuxième catégorie comblent cette lacune en proposant de détecter une plus grande palette d'évènements et en leur donnant un sémantique plus forte.

Notre approche permet de détecter plusieurs évènements de foule. Notre contribution s'affiche dans l'utilisation du modèle directionnel pour suivre un groupe de personnes ayant la

même direction de mouvement plutôt que de détecter et suivre chaque personne individuellement. Ceci permet de modéliser plus naturellement des évènements de foule comme la séparation de deux groupes ou la fusion de deux groupes.

Finalement, nous avons abordé les approches d'estimation des flux. Nous avons montré qu'il y avait des méthodes variées et qui donnaient des résultats convaincants. Cependant, ces approches ont deux problèmes majeurs, soit elles sont trop gourmandes en calculs (approches stéréoscopiques), soit elles sont applicables à une configuration spécifique et nécessitent un effort conséquent si on veut changer de point de vue ou la position de la caméra. Nous pensons que les approches spatiotemporelles sont celles qui offrent le meilleur compromis entre temps de calcul, précision et permissivité par rapport au changement de configuration. Néanmoins, les approches de l'état de l'art présentent une limite notable ; elles ne s'exécutent pas en temps réel et ont besoin d'un certain délai avant de donner un résultat de comptage (le temps de construction de la carte spatiotemporelle).

Notre approche permet de compter le nombre d'individus franchissant une ligne de comptage à partir de vidéos issues de caméras monoculaires. Elle fait partie des approches spatiotemporelles avec l'originalité d'éviter la détection des individus qui s'arrêtent sur la ligne de comptage en utilisant les vecteurs de flux optique. Ces vecteurs permettent également d'obtenir l'orientation des personnes de manière plus précise. Notre approche est aussi capable de détecter en temps réel les blobs qui franchissent la ligne de comptage contrairement aux approches précédentes qui attendent la construction complète de la carte spatiotemporelle avant d'entamer la détection des blobs. L'utilisation d'un modèle de régression linéaire pour chaque angle de prise de vue permet à notre approche d'être indépendante de l'angle de prise de vue.

Dans les chapitres suivants, nous présentons nos contributions dans les domaines de l'analyse du comportement humain dans les scènes individuelles et les scènes de foule. Le chapitre suivant introduit les descripteurs de niveau intermédiaire développés durant la thèse et qui sont utilisés dans la plupart de nos contributions applicatives. Nous mettons en application ces descripteurs dans un chapitre consacré à la reconnaissance d'actions et dans un autre chapitre consacré à l'analyse des scènes de foule.

# Chapitre 3

# Descripteurs pour l'analyse du comportement humain

## 3.1 Introduction

Nous exposons dans ce chapitre nos principales contributions en termes de descripteurs de niveau intermédiaire pour l'analyse du comportement humain. Nous présentons dans un premier temps le modèle de magnitude et le modèle directionnel. Ces deux modèles estiment les vitesses et les orientations de mouvement dominantes dans chaque région de la scène. Le modèle directionnel permet notamment de capturer les modalités de mouvement dans les scènes structurées et non structurées.

Dans un deuxième temps, nous présentons deux ensembles de descripteurs pour la dynamique de groupes de personnes. Le premier ensemble de descripteurs est estimé depuis des images en grille où l'on suit des groupes de personnes qui se déplacent dans la même direction. Ces descripteurs permettent de développer des détecteurs d'évènements sémantiques. Nous proposons une approche reposant sur le modèle directionnel pour détecter et suivre des groupes de personnes en temps réel. Cette approche est également applicable aux scènes de foule car elle ne nécessite pas l'extraction des silhouettes des personnes.

Nous présentons ensuite un deuxième ensemble de descripteurs caractérisant les groupes de personnes franchissant une ligne virtuelle. Nous proposons une approche qui permet d'estimer en temps réel les descripteurs liés à un groupe de personnes dès que ces derniers franchissent la ligne virtuelle. Ceci permet d'estimer le nombre de personnes franchissant la ligne virtuelle indépendamment de la position de la caméra.

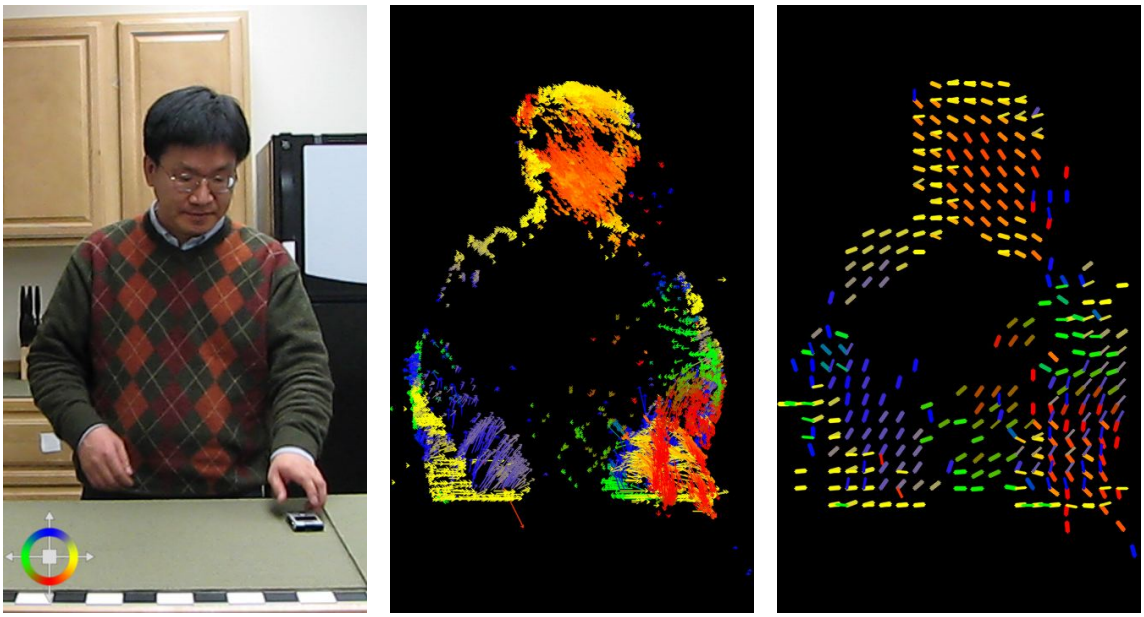
## 3.2 Modèle directionnel et modèle de magnitude

Nous introduisons deux descripteurs de niveau intermédiaire qui permettent de modéliser le comportement des personnes dans la vidéo en termes d'orientation et de vitesse de mouvement.

Après avoir calculé les vecteurs de mouvement, l'étape suivante consiste à diviser la scène en une grille de $M \times N$ blocs. Puis, chaque vecteur de mouvement est associé au bloc qui lui

correspond selon son origine. La taille des blocs influe sur la précision du système et sera étudiée dans la Section 4.4.3.

Un algorithme de regroupement de données circulaires est ensuite appliqué aux orientations des vecteurs de flux optique dans chaque bloc. L'ensemble des $M \times N$ distributions circulaires associées est appelé "modèle directionnel". La Figure 3.1 montre la construction d'un modèle directionnel associé à l'action 'answerPhone'.



(a) (b) (c)

FIGURE 3.1 – Modèle directionnel pour l'action 'answerPhone'. (a) image courante, (b) vecteurs de flux optique, (c) modèle directionnel associé à la séquence vidéo

Dans ce travail, nous regroupons les données circulaires en utilisant un mélange de lois von Mises. Ainsi, la probabilité d'obtenir une orientation $\theta$ par rapport à un bloc $B_{x,y}$ est définie par la formule suivante :

$$p_{x,y}(\theta) = \sum_{i=1}^{K} \psi_{i_{x,y}} \cdot V\left(\theta; \phi_{i_{x,y}}, \gamma_{i_{x,y}}\right) \tag{3.1}$$

où :

– $K$ représente le nombre de lois du mélange. Nous avons choisi empiriquement $K = 4$, pour correspondre aux quatre points cardinaux.

- $\psi_{i_{x,y}}$, $\phi_{i_{x,y}}$, $\gamma_{i_{x,y}}$ sont respectivement le poids, l'angle moyen et le paramètre de concentration de la $i^{\text{ème}}$ distribution du bloc $B_{x,y}$.
- $V(\theta;\phi,\gamma)$ est la loi de von Mises de direction $\phi$ avec un paramètre de concentration $\gamma$. Elle possède la fonction de densité de probabilité suivante sur l'intervalle $[0,2\pi[$ :

$$V(\theta;\phi,\gamma) = \frac{1}{2\pi I_0(\gamma)} \exp[\gamma \cos(\theta-\phi)] \tag{3.2}$$

où $I_0(\gamma)$ est la fonction de Bessel modifiée de première espèce d'ordre 0 définie par :

$$I_0(\gamma) = \sum_{r=0}^{\infty} \left(\frac{1}{r!}\right)^2 \left(\frac{1}{2}\gamma\right)^{2r} \tag{3.3}$$

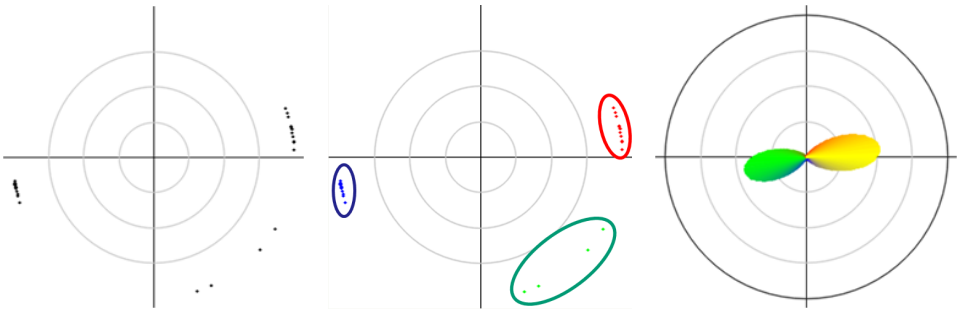
Par analogie, nous regroupons les magnitudes des vecteurs du flux optique dans chaque bloc grâce à des mélanges gaussiens. L'ensemble des mélanges gaussiens estimés représente le modèle de magnitude. Ainsi, la probabilité d'une magnitude $v$ par rapport au bloc $B_{x,y}$ est définie de la façon suivante :

$$p_{x,y}(v) = \sum_{i=1}^{J} \omega_{i_{x,y}} G(v;\mu_{i_{x,y}},\sigma^2_{i_{x,y}}) \tag{3.4}$$

où :

- $\omega_{i_{x,y}}, \mu_{i_{x,y}}, \sigma^2_{i_{x,y}}$ sont respectivement le poids, la moyenne et la variance de la $i^{\text{ème}}$ Gaussienne.
- $J$ est le nombre de Gaussiennes ($J = 4$ dans nos expérimentations).

Pour chaque image, nous mettons à jour les paramètres des mélanges Gaussiens grâce à une approximation de k-moyennes décrite dans [KB01]. Nous l'utilisons également pour estimer les paramètres des mélange de lois de von Mises en adaptant l'algorithme afin de gérer les données circulaires et en prenant en compte l'inverse de la variance en tant que paramètre de dispersion ($\gamma = 1/\sigma^2$).



(a) Données d'entrée (b) Groupes détectés (c) Représentation de la densité de probabilité sur le cercle unitaire

FIGURE 3.2 – Illustration des résultats de notre algorithme de regroupement de données circulaires

L'algorithme 1 liste les principales étapes l'algorithme de regroupement de données circulaires avec un mélange de lois de von Mises. La Figure 3.2 montre le résultat du regroupement d'un ensemble de données circulaires à l'aide de notre approche.

Nous notons dorénavant le modèle de la séquence $s$ par $Sm(s) = (Dm(s), Mm(s))$, où $Dm(s)$ et $Mm(s)$ sont respectivement le modèle directionnel et le modèle de magnitude associés à la séquence $s$. La Figure 3.3 montre les modèles directionnels et de magnitude de quelques séquences vidéos issues de la base KTH.

Si on prend une séquence vidéo dans laquelle se trouve une personne effectuant une action, le modèle de cette séquence est considéré comme la signature de l'action effectuée. La Section 4 propose notre approche de reconnaissance des actions humaines à partir du modèle d'une séquence.

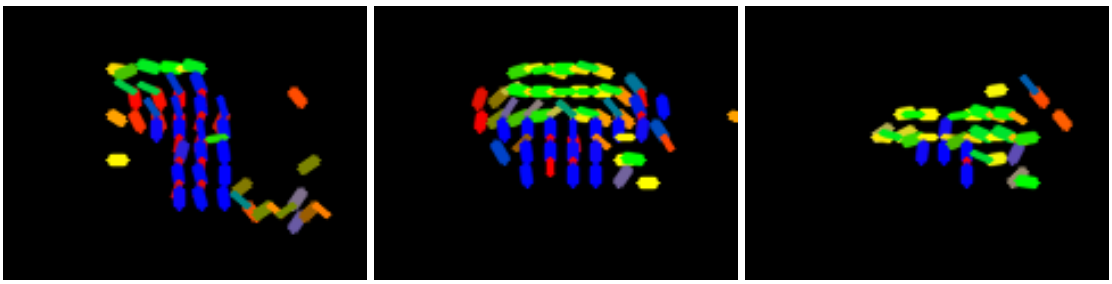
Le modèle directionnel est également employé pour analyser le comportement humain dans une scène de foule. Nous montrons dans le Chapitre 5.1 une approche d'extraction des motifs de mouvement en regroupent les blocs voisins du modèle directionnel.

Dans ce qui suit, nous introduisons deux approches de détection de groupes de personnes dans une scène de foule.

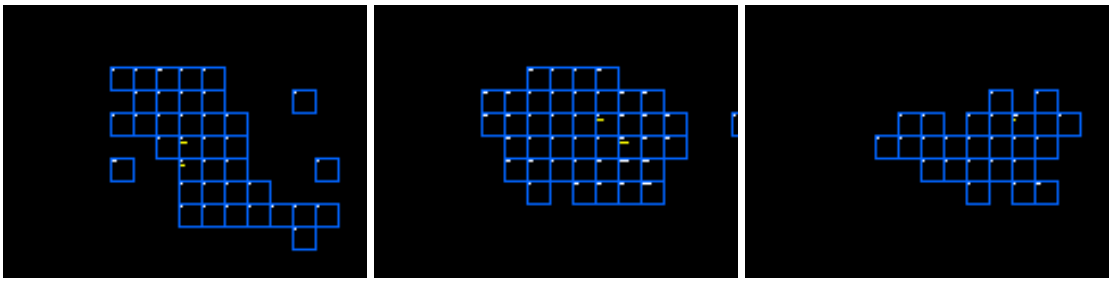
**Algorithme 1** Algorithme d'estimation des paramètres d'un mélange de lois de von Mises

**Entrée :** -un élément $x$ de $\mathbb{R}$ qui représente une orientation de mouvement
-un mélange de $K$ lois de von Mises
**Sortie :** paramètres du mélange et groupes mis à jour
initialiser le taux d'apprentissage $\alpha$ = 1/400
initialiser le seuil de mise en correspondance $\beta$ = 2.5
$c \leftarrow 0$
**pour** $i = 1$ à $K$ **faire**
    {Recherche de la première loi qui convient à la donnée}
    **si** $c = 0$ and $x - \theta_i \leq \beta^2/\gamma_i$ **alors**
        $c \leftarrow i$
    **fin si**
**fin pour**
**si** $c \neq 0$ **alors**
    {Si on trouve cette loi, on met à jour les paramètres}
    **pour** $i = 1$ to $K$ **faire**
        $\psi_i \leftarrow \psi_i(1-\alpha)$
    **fin pour**
    $\psi_c \leftarrow \psi_c + \alpha$
    $\rho \leftarrow \alpha\psi_c(x-\theta_c)$
    $\theta_c \leftarrow \theta_c + \rho$
    $\gamma_c \leftarrow (\gamma_c^{-1} + \rho^2 - \gamma_c^{-1})^{-1}$
    $n_c \leftarrow n_c + 1$
**sinon**
    {La donnée ne peut être mise en correspondance avec aucune loi}
    $n_k \leftarrow 1$
    $\theta_k \leftarrow x$
    $\gamma_k \leftarrow \gamma_0$
    **pour** $i = 1$ to $K$ **faire**
        $\psi_i \leftarrow \dfrac{n_i}{\sum_{j=1}^{k} n_j}$
    **fin pour**
**fin si**
trier les lois du mélange selon la valeur de $\psi \times \gamma$

(a) Échantillon d'images



(b) Modèle directionnel



(c) Modèle de magnitude

FIGURE 3.3 – Échantillon d'images avec les modèles de direction et de magnitude qui leur sont associés

## 3.3 Descripteurs sur les groupes de personnes en mouvement

Nous présentons dans cette section deux nouvelles approches de détection de groupes de personnes dans des scènes de foules. La première approche traite la scène comme des cartes spatiotemporelles estimées grâce à une ligne prédéfinie. Elle permet d'extraire des descripteurs permettant d'estimer efficacement le nombre de personnes traversant une ligne virtuelle. La seconde approche traite la scène comme des images en grille. Elle permet d'en extraire des descripteurs pour la détection d'évènements sémantiques.

Dans ce qui suit, nous utilisons le terme *blob* pour désigner une région de pixels connectés dans une image. C'est l'abréviation de sa définition anglaise : *Binary Large OBject* qui signifie large objet binaire en français. Nous l'employons fréquemment dans la Section suivante pour désigner un objet franchissant la ligne virtuelle.

### 3.3.1 Extraction des blobs depuis une carte spatiotemporelle 2D

Tout d'abord, nous définissons une ligne virtuelle comme expliqué dans la Section 2.2.1. L'étape suivante consiste à calculer les vecteurs du flux optique des pixels appartenant à la ligne de comptage pour chaque image de la séquence vidéo. Pour cela, nous utilisons la méthode décrite dans la Section 2.3.1. L'orientation et la vitesse des vecteurs obtenus sont empilées à travers le temps afin de construire des cartes spatiotemporelles d'orientation et de vitesse.

Ces deux cartes sont utilisées lors de l'étape de détection en ligne des blobs pour détecter et mettre à jour leurs caractéristiques (position, vitesse, orientation et taille) jusqu'à ce que les blobs prennent une forme définitive (i.e. que le ou les personnes aient complètement franchi la ligne de comptage).

Nous utilisons les notations suivantes :

- $L$ est la ligne de comptage de longueur $l$.
- $p_{i,t}$ est un pixel de la carte spatiotemporelle $Map_L$ construite à partir de la ligne $L$ à l'instant $t$ avec $i = \overline{0..l-1}$.

– $OF(p_{i,t})$ est le vecteur de flux optique ayant pour origine le pixel $p_{i,t}$. Chaque vecteur est défini par sa vitesse et son orientation comme suit :

$$OF(p_{i,t}) = \begin{pmatrix} vitesse(p_{i,t}) \\ orientation(p_{i,t}) \end{pmatrix} \tag{3.5}$$

Les cartes spatiotemporelles de mouvements et d'orientations sont obtenues grâce au calcul du flux optique sur la ligne de comptage. Le passage d'une ou d'un groupe de personnes sur la ligne virtuelle forme un blob sur la chaque carte spatiotemporelle. Nous devons donc détecter ces blobs. Avant de décrire l'algorithme de détection des blobs, nous présentons les descripteurs modélisant un blob de façon formelle.

Un blob identifié par $Id$ est défini par le vecteur suivant :

$$B(Id) = (P, N, \alpha, \beta, l, h, O, V) \tag{3.6}$$

où :

– $P$ est l'ensemble de pixels du blob et $N$ leurs nombre.
– $l$ et $h$ sont respectivement la largeur et la hauteur du rectangle minimum englobant les pixels du blob.
– $\alpha$ et $\beta$ sont les coordonnées du coin supérieur gauche de ce rectangle.
– $O$ correspond à l'orientation moyenne du blob et $V$ à sa vitesse moyenne.

Afin de simplifier l'écriture des équations, nous notons ces éléments de la façon suivante : $B(Id).P$, $B(Id).N$, $B(Id).l$, etc.

L'algorithme de détection des blobs (voir Algorithme 2) est utilisé après le calcul des vecteurs du flux optique des pixels de la ligne de comptage pour l'image courante. Son but est de mettre à jour en temps réel un ensemble des blobs qu'on note $S$.

Notre algorithme prend uniquement en compte les pixels dont le flux optique a une vitesse non nulle pour filtrer les personnes statiques. Les pixels voisins de la colonne courante dans la carte spatiotemporelle ayant une orientation similaire sont regroupés dans le même blob. En

d'autres termes, deux pixels $p_{i,t}$ et $p_{j,s}$ sont considérés comme voisins si $i \in \{j, j-1, j+1\}$ et $t \in \{s, s-1, s+1\}$.Deux orientations sont considérées similaires si leur différence est inférieure à $\pi/2$ radians. Nous avons choisi cette fourchette pour palier à certaines imprécisions du flux optique car il est difficile d'avoir deux vecteurs dont les orientations sont exactement égales.

**Algorithme 2** Détection de blob en ligne pour une image t

Entrée : - Ligne de comptage $L$
- Pixels $p_{i,q} \in Map_L$
- Ensemble de blobs $S = \{B(1), B(2), \cdots, B(m)\}$
**retour** Ensemble de blobs mis à jour $S$
**pour** $i \leftarrow 0$ to $l-1$ **faire**
  **si** $velocity(p_{i,t}) \neq 0$ **alors**
    *//Grouper les pixels voisins sur la même ligne*
    **si** $(p_{i-1,t} \in B(j))$ et ($B(j).O$ similaires à $orientation(p_{i,t})$) **alors**
      $B(j).P \Leftarrow B(j).P \cup \{p_{i,t}\}$
      Mettre à jour $B(j)$
    **fin si**
    *//Grouper les pixels sur des lignes différentes*
    **si** $orientation(p_{i,t})$ similaire à $orientation(p_{i,t-1})$ **alors**
      **si** $(p_{i,t} \in B(j))and(p_{i,t-1} \in B(k))$ **alors**
        *//Grouper les blobs voisins*
        $B(j).P \Leftarrow B(j).P \cup B(k).P$
        Supprimer $B(k)$
        Mettre à jour $B(j)$
      **sinon si** $p_{i,t-1} \in B(k)$ **alors**
        $B(k).P \Leftarrow B(k).P \cup p_{i,t}$
        Mettre à jour $B(k)$
      **fin si**
    **fin si**
    *//Mettre un pixel non-groupé dans un nouveau blob*
    **si** $p_{i,t} \notin S$ **alors**
      Créer un nouveau blob $B(m+1)$
      Initialiser $B(m+1)$ avec le pixel $p_{i,t}$
      $S \leftarrow S \cup \{B(m+1)\}$
      $m \leftarrow m+1$
    **fin si**
  **fin si**
**fin pour**

Les blobs obtenus à partir de la colonne courante sont regroupés avec ceux obtenus à partir des itérations précédentes de l'algorithme :

– deux blobs $B(a)$ et $B(c)$ sont regroupés à condition d'être voisins et d'avoir des orientations similaires. Le blob $B(a)$ est voisin du blob $B(c)$ s'il existe un pixel $p_{j,s}$ dans $B(a).P$ qui est lui-même voisin d'un pixel $p_{k,r}$ dans $B(c).P$.
– Lorsqu'un nouveau pixel $p_{i,t}$ est ajouté au blob $B(Id)$, alors l'algorithme met les paramètres du blob à jour par le biais des formules suivantes :

$$B(Id).P = B(Id).P \cup \{p_{i,t}\} \tag{3.7}$$

$$B(Id).N = B(Id).N + 1 \tag{3.8}$$

$$B(Id).V = \frac{\sum_{p \in B(id).P} vitesse(p)}{B(Id).N} \tag{3.9}$$

$$B(Id).O = atan2\left(\frac{\sum_{p \in B(id).P} \sin(orientation(p))}{\sum_{p \in B(id).P} \cos(orientation(p))}\right) \tag{3.10}$$

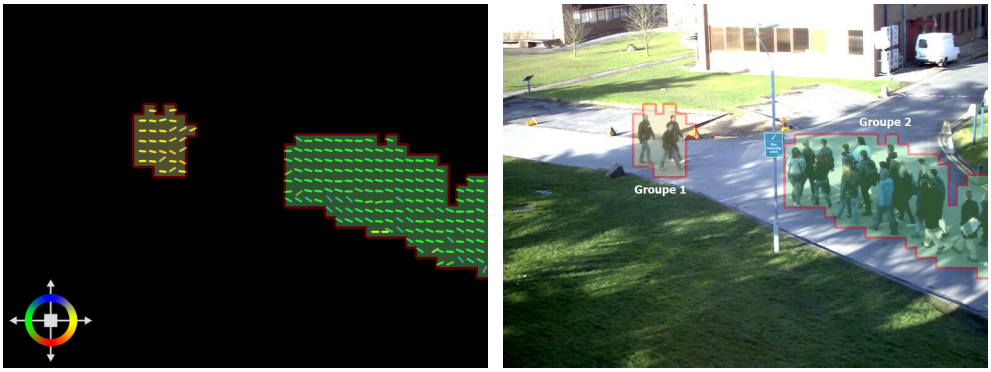
L'algorithme sauvegarde les blobs de sorte qu'ils puissent être mis à jour lors des itérations suivantes. Une itération représente un empilement d'une image dans les cartes spatiotemporelles. Si un blob n'a pas été mis à jour après deux itérations, alors ce blob est considéré comme une entité ayant franchi la ligne virtuelle entièrement. L'étape d'extraction des descripteurs des blobs prend uniquement en compte ceux qui ont franchi la ligne de comptage et leur assigne les caractéristiques décrites dans la Formule 3.6. Nous détaillons dans la Section 5.3, une approche permettant d'estimer le nombre de personnes dans un blob ainsi que leurs direction.

### 3.3.2 Groupes de personnes dans une représentation sous forme de grille

#### Groupement des blocs

L'objectif de cette étape consiste à regrouper des blocs d'images pour obtenir les groupes représentant des entités ayant une orientation et une vitesse similaires. Ces groupes peuvent inclure une ou plusieurs personnes. Les critères d'attribution d'un bloc à un groupe sont l'orientation du mouvement et la vitesse. Ainsi, les blocs voisins ayant une magnitude moyenne et une orientation principale similaires feront partie du même groupe.

Chaque bloc $B_{x,y}$ est défini par sa position $P_{x,y} = (x,y); x = \overline{1..Bx}, y = \overline{1..By}$ et son orientation $\Omega_{x,y} = \mu_{0,x,y}$. L'image 3.4 montre les résultats obtenus lors de cette étape.




(a) Segmentation du bloc (b) Image initiale et groupes résultants

FIGURE 3.4 – Regroupement des blocs dans une image

### Suivi de groupes

Le suivi de groupes permet de reconstituer la trajectoire d'un groupe de personnes, de son apparition dans le champ de la caméra jusqu'à sa disparition. Cette opération est réalisée en établissant une concordance entre les barycentres des groupes d'une image $f$ avec ceux de l'image suivante $f+1$. Chaque image $f$ est définie par ses groupes $\{C_{1,f}, C_{2,f}, ..., C_{n_f,f}\}$ où $n_f$ est le nombre de groupes détectés dans l'image $f$. Chaque groupe $C_{i,f}$ est décrit par son barycentre $O_{i,f}$ et son orientation moyenne $X_{i,f}$. Le groupe $C_{m,f+1}$ est mis en correspondance avec le groupe $C_{i,f}$ si son barycentre est le plus proche de $C_{i,f}$ sans excéder une distance minimale. Formellement, il doit respecter les deux conditions suivantes :

$$\begin{cases} m = \underset{j}{\operatorname{argmin}}(D(O_{i,f}, O_{j,f+1})) \\ et \\ D(O_{i,f}, O_{m,f+1}) < \tau \end{cases} \tag{3.11}$$

où $\tau$ est la distance maximale entre deux groupes mis en correspondance (nous choisissons empiriquement $\tau = 5$). S'il n'existe aucune correspondance pour le groupe $C_{i,f}$ (c'est-à-dire s'il

n'existe aucun groupe $C_{m,f+1}$ qui remplit ces deux conditions), alors ce groupe disparaît et n'est plus suivi dans les prochaines images.

## 3.4 Conclusion

Nous avons introduit dans ce chapitre des descripteurs de niveau intermédiaire pour l'analyse du comportement humain. Ces descripteurs sont basés sur le mouvement comme information de bas niveau.

Nous avons d'abord introduit le modèle directionnel et le modèle de magnitude dans la Section 3.2. Ce sont des grilles où chaque cellule modélise les orientations et vitesses de mouvement majeures de la zone de la scène couverte par une cellule. Nous utilisons pour cela des mélanges de loi de von Mises (orientation) et des mélanges de gaussiennes (magnitude). Nous exposons dans le Chapitre 4 une application directe du modèle directionnel et de magnitude pour la reconnaissance d'actions dans une scène individuelle. Nous montrons dans la première section du Chapitre 5 que le modèle directionnel permet d'extraire les différents motifs de mouvements dans les scènes structurées et non structurées.

Nous avons ensuite introduit deux approches de détection des groupes de personnes en temps réel dans des scènes de foule. Elles utilisent l'orientation du mouvement et le voisinage spatial sur le plan de l'image comme critère de regroupement.

La première approche (Section 3.3.1) permet de détecter des blobs traversant une ligne virtuelle au moyen de cartes spatiotemporelles. Elle se base sur le regroupement des pixels voisins sur la ligne de comptage ayant une orientation similaire. Cette approche a plusieurs avantages car elle utilise le mouvement et le regroupement en temps réel des blobs. Premièrement, elle permet de connaitre directement la direction d'un blob lorsqu'il passe sur la ligne virtuelle. Deuxièmement, elle ignore les objets qui sont immobiles sur la ligne virtuelle car leur vitesse de mouvement est nulle, contrairement aux approches spatiotemporelles de l'état l'art 2.6.3. Enfin, nous soulignons le fait que notre approche puisse détecter des blobs en temps réel contrairement aux autres approches spatiotemporelles.

La deuxième approche (Section 3.3.2) se base sur une représentation du mouvement en grilles afin de regrouper les cellules voisines qui ont une orientation similaire. Cette méthode permet de détecter et de regrouper des objets en mouvement sans se soucier de leur morphologie. Ceci est notamment utile dans les scènes de foule où les occlusions ne permettent pas de distinguer toutes les parties du corps d'une personne. Par rapport à une soustraction d'arrière-plan, notre méthode a l'avantage de différencier les groupes selon leur direction. Ceci permet de détecter certains types d'évènements tels que la séparation ou la fusion de groupes.

Ces descripteurs peuvent être utilisés pour traiter plusieurs problèmes. Nous focalisons ce mémoire de thèse sur les applications de l'analyse du comportement humain dans un scène de foule et une scène individuelle comme nous l'avons décrit dans nos objectifs (Section 1.3). Nous listons chaque problème ainsi que la section ou le chapitre qui le traite :

1. La reconnaissance des actions humaines (Chapitre 4)
2. L'extraction des motifs de mouvement (Section 5.1)
3. La détection d'évènements de foule (Section 5.2)
4. L'estimation des flux (Section 5.3)

Le chapitre suivant aborde le premier objectif qui est la reconnaissance des actions humaines.

# Chapitre 4

# Analyse du comportement humain dans des scènes individuelles

## 4.1 Introduction

La reconnaissance d'actions humaines est un sujet particulièrement complexe dans le domaine de la vision par ordinateur. Cela consiste en la classification automatique des actions réalisées par un individu dans une séquence vidéo. La reconnaissance des actions est cruciale dans de nombreux domaines comme la vidéosurveillance, l'interaction homme-machine et l'indexation des vidéos.

Dans la littérature, nous trouvons certaines approches qui détectent les actions à partir d'images fixes, tandis que d'autres ont recours à des vidéos monoculaires, des vidéos stéréoscopiques ou à des maillages 3D. Nous avons traité, dans notre travail, les séquences vidéo enregistrées par des caméras monoculaires. Dans ce type de vidéos, la détection d'actions peut se faire en combinant des informations spatiales et temporelles [JBEJ94]. Cet intérêt pour les vidéos monoculaires au détriment des vidéos stéréoscopiques résulte du fait qu'elles sont couramment utilisées, moins gourmandes en ressources réseaux et informatiques et plus économiques.

Nous proposons une approche de reconnaissance des actions humaines à l'aide du modèle directionnel et modèle de magnitude introduits en Section 3.2. La Section suivante décrit les étapes de notre approche dans laquelle nous spécifions le niveau bas, intermédiaire et sémantique. La Section 4.3 présente une mesure de distance qui nous permet de reconnaitre l'action d'une vidéo à partir d'une base d'exemples. Nous expérimentons et commentons les résultats de notre approche dans la Section 4.4. Nous concluons ce chapitre dans la Section 4.5.

## 4.2 Description de l'approche

Pour reconnaitre les actions réalisées par une personne, nous proposons une approche qui suit la méthodologie en trois niveaux décrite dans la Section 1.4. La Figure 4.1 distingue les différentes informations calculées à chaque niveau de notre approche :

– Bas niveau : cette étape a pour but de quantifier le mouvement à partir des vecteurs de flux optique, afin d'estimer le modèle directionnel et le modèle de magnitude. Nous appli-

quons d'abord le détecteur de points d'intérêt vu dans la Section 2.2.3. Nous appliquons aux points d'intérêt la méthode d'estimation des vecteurs du flux optique vue dans la Section 2.3.1.

– Niveau intermédiaire : cette étape estime le modèle directionnel et le modèle de magnitude pour l'intégralité de la séquence. Ces modèles serviront à constituer le modèle correspondant à l'action contenue dans la séquence. L'étape d'estimation du modèle directionnel et de magnitude a été présentée dans la Section 3.2.

– Niveau sémantique : cette étape a pour but de reconnaitre l'action dans une vidéo en comparant ses modèles d'orientation et de magnitude estimés dans l'étape intermédiaire à partir des séquences vidéos de référence à l'aide d'une mesure de distance. Nous détaillons cette étape dans la section suivante.

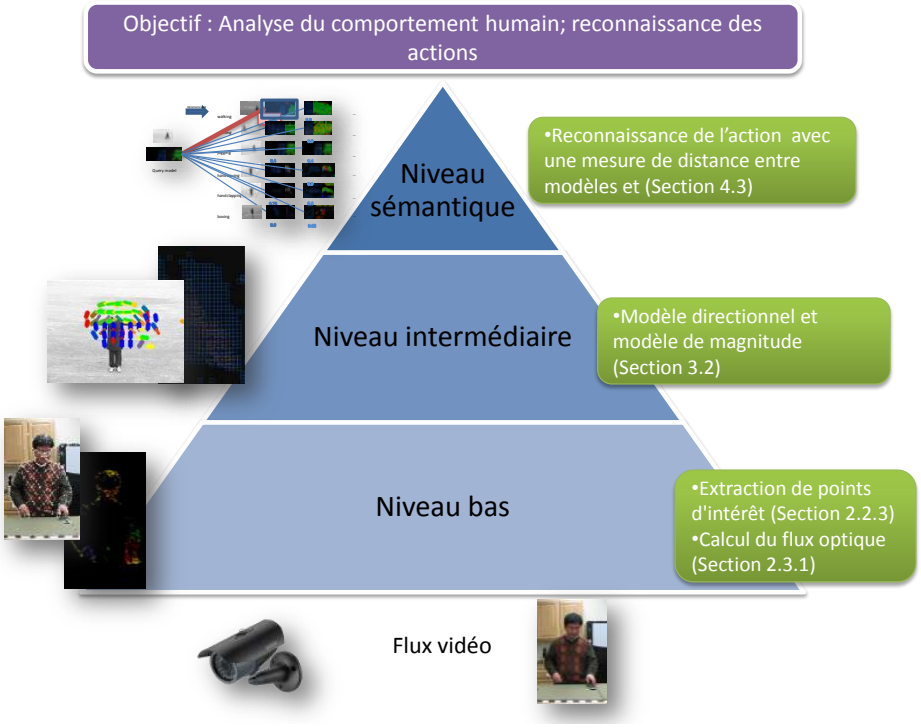



FIGURE 4.1 – Schéma de l'approche de reconnaissance d'actions

## 4.3 Reconnaissance de l'action (niveau sémantique)

Une fois le modèle de la séquence vidéo calculé (comme indiqué dans la Section 3.2), nous détectons l'action correspondant à cette séquence en fonction des vidéos de référence. Les actions sont détectées en comparant le modèle d'une séquence avec les modèles associés aux séquences de référence en utilisant une mesure de distance. L'action associée au modèle ayant la distance la plus petite par rapport au modèle d'une séquence est retenue. La Figure 4.2 illustre ce processus où nous calculons d'abord les distances entre le modèle requête à gauche et les modèles de référence à droite. Dans ce cas, l'évènement retenu est l'évènement 'walking' car une vidéo de référence correspondant à l'évènement 'walking' donne la plus petite distance (qui est de 0.02) avec le modèle de référence.

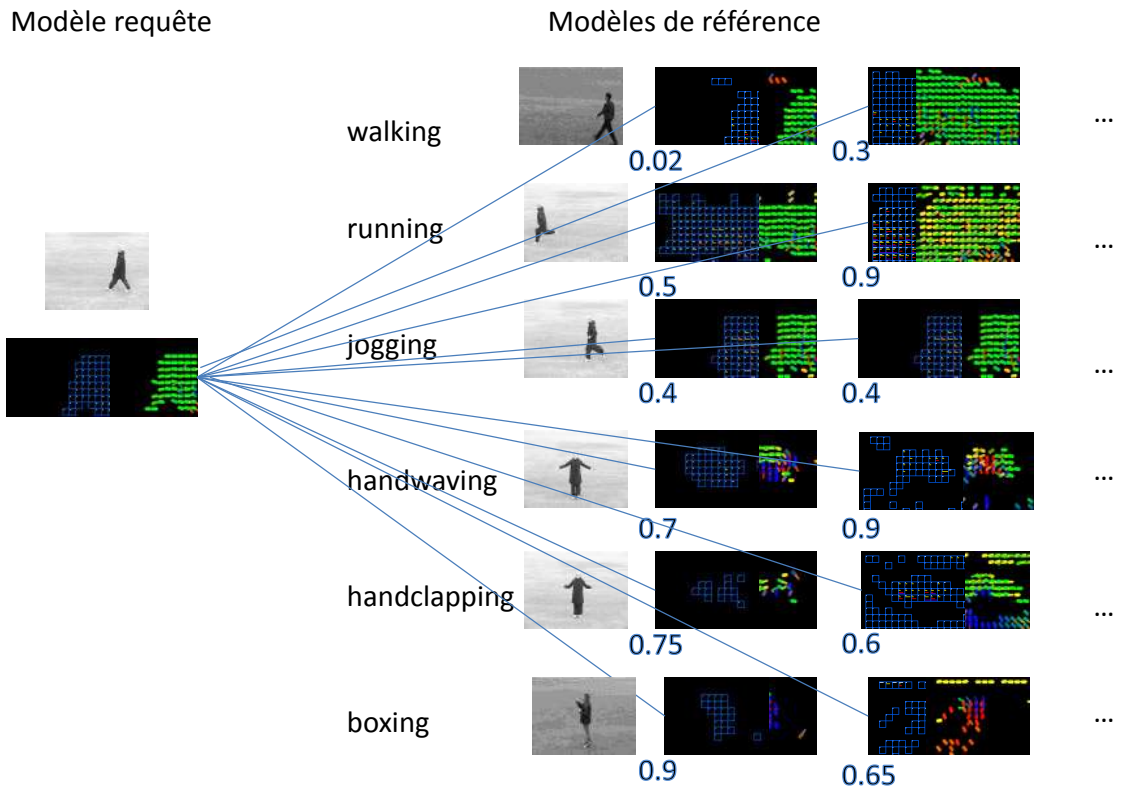



FIGURE 4.2 – Illustration de notre processus de reconnaissance des actions en choisissant la plus petite distance entre le modèle requête et les modèles de référence

Soient $T = \{t_1, t_2, ..., t_n\}$ un ensemble de $n$ séquences vidéos avec leurs modèles respectifs $\{Sm(t_1), Sm(t_2), ..., Sm(t_n)\}$ et $q$ une séquence requête avec son modèle $Sm(q)$. La distance entre $Sm(q)$ et une séquence de référence $Sm(t_l)$ est définie par :

$$D(Sm(q), Sm(t_l)) = Norm(A_{Dm(q),Dm(t_l)}) + Norm(B_{Mm(q),Mm(t_l)}) \tag{4.1}$$

où :

- $Norm$ correspond à la norme L2.
- Les matrices $A_{Dm(q),Dm(t_l)}$ et $B_{Mm(q),Mm(t_l)}$ de dimensions $W \times H$ contiennent, respectivement, les distances entre chaque élément des deux modèles directionnels $Dm(q)$ et $Dm(t_l)$ et des deux modèles de magnitude $Mm(q)$ et $Mm(t_l)$
- Chaque élément $A_{M,M'}(x,y)$ est défini par la formule suivante :

$$A_{M,M'}(x,y) = \sum_{i=1}^{K} \left( \psi_{i_{x,y}} \psi'_{i_{x,y}} Dist_d(V_{i_{x,y}}, V'_{i_{x,y}}) \right) \tag{4.2}$$

où :

- $\psi_{i_{x,y}}$ (resp. $\psi'_{i_{x,y}}$) et $V_{i_{x,y}}$ (resp. $V'_{i_{x,y}}$) représentent le poids et la variance de la $i^{\text{ème}}$ loi de von Mises associés au modèle directionnel $M$ (resp. $M'$) dans le bloc $B_{x,y}$.
- $Dist_d(V,V')$ est la distance de Bhattacharyya entre les deux lois de von Mises $V$ et $V'$ définie par l'équation suivante :

$$Dist_d(V,V') = \sqrt{1 - \int_{-\infty}^{+\infty} \sqrt{V(\theta)V'(\theta)} d\theta} \tag{4.3}$$

où $Dist_d(V,V')$ est comprise entre 0 et 1. Cette équation peut être calculée grâce à cette solution de forme fermée :

$$Dist_d(V,V') = \sqrt{1 - \sqrt{\frac{1}{I_0(\gamma)I_0(\gamma')}} I_0\left( \frac{\sqrt{\gamma^2 + \gamma'^2 + 2\gamma\gamma' cos(\phi - \phi')}}{2} \right)} \tag{4.4}$$

où $\phi$ (resp. $\phi'$) et $\gamma$ (resp. $\gamma'$) sont respectivement l'angle moyen et le paramètre de dispersion de la distribution $V$ (resp. $V'$).

Par analogie, nous définissons chaque élément $B_{N,N'}(x,y)$ par l'équation suivante :

$$B_{N,N'}(x,y) = \sum_{i=1}^{K} \left( \omega_{i_{x,y}} \omega'_{i_{x,y}} Dist_m(G_{i_{x,y}}, G'_{i_{x,y}}) \right) \tag{4.5}$$

où

- $\omega_{i_{x,y}}$ (resp. $\omega'_{i_{x,y}}$) et $G_{i_{x,y}}$ (resp. $G'_{i_{x,y}}$) représentent le poids de la $i^{\text{ème}}$ gaussienne associée au modèle de magnitude $N$ (resp. $N'$) dans le bloc $B_{x,y}$.
- $Dist_m(G,G')$ est la distance de Bhattacharyya entre deux gaussiennes $G$ et $G'$ définies dans la solution de forme fermée suivante :

$$Dist_m(G,G') = \frac{(\mu - \mu')^2}{4(\sigma^2 + \sigma'^2)} + \frac{1}{2} \ln \left( \frac{\sigma^2 + \sigma'^2}{2\sigma\sigma'} \right) \tag{4.6}$$

où $\mu$ (resp. $\mu'$) et $\sigma^2$ (resp. $\sigma'^2$) sont respectivement la moyenne et la variance de la gaussienne $G$ (resp. $G'$).

## 4.4 Expérimentations et résultats

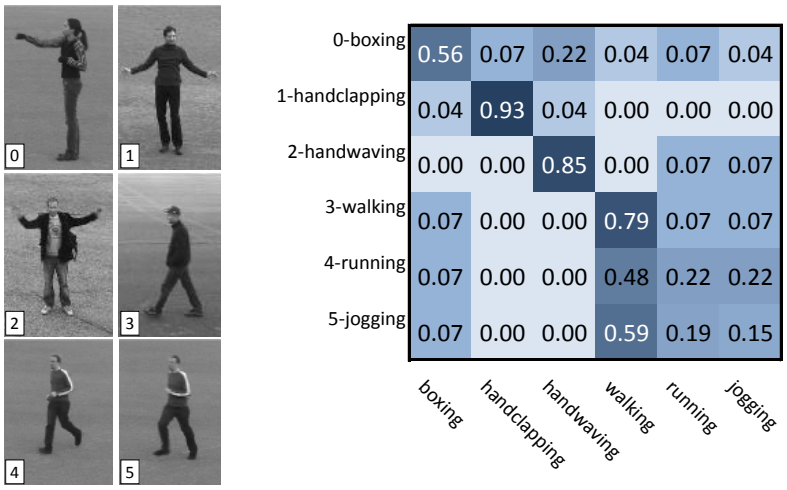
Nous montrons dans cette section l'efficacité de notre approche sur des vidéos qui contiennent une variété d'actions quotidiennes issues des bases de vidéo KTH et ADL. Nous présentons les résultats de notre approche à l'aide de matrices de confusion où les données horizontales correspondent à la vérité terrain et les données verticales correspondent à nos résultats. Nous présentons ensuite un tableau comparatif suivi d'une étude sur les effets de la taille des blocs et du nombre de classes d'actions sur les performances du système.

### 4.4.1 Efficacité de la reconnaissance des actions

*Base vidéo KTH [LL04] :*C'est une base de vidéos de faible résolution (images en niveau de gris de $160 \times 120$ pixels) regroupant 6 actions effectuées plusieurs fois par 25 personnes. Cette base contient des vidéos en environnement intérieur et extérieur. Nous divisons l'ensemble de

données en deux ensembles, comme suggéré par Schuldt et al. [SLC04] : un ensemble d'apprentissage qui contient les séquences de référence (16 personnes) et un ensemble de test qui contient les séquences requête (9 personnes). L'ensemble d'apprentissage comporte les personnes 'person01' à 'person16' et l'ensemble de test comporte les personnes 'person17' à 'person25'. Nous utilisons des blocs de taille de $5 \times 5$.

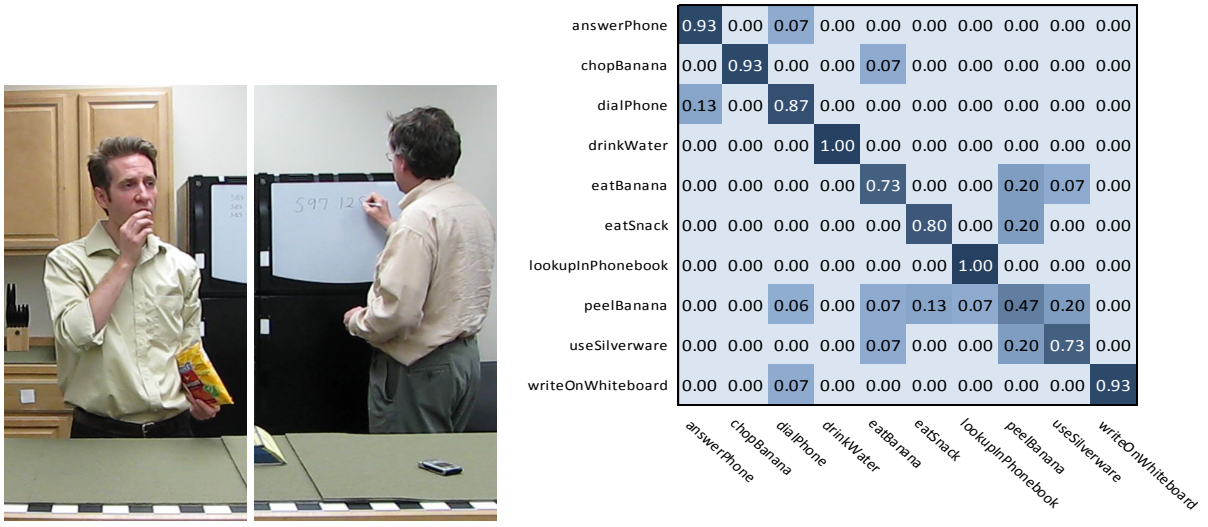
Quelques exemples d'actions ainsi que la matrice de confusion sont présentés dans la Figure 4.3. Notre approche aboutit à de meilleurs résultats pour les trois premières actions de la base de vidéos lorsque la personne est immobile. En revanche, notre système assimile les actions 'run' et 'jogging' à l'action 'walk'. Ceci est dû au fait que ces actions diffèrent légèrement de par la vitesse et la longueur de la foulée tout en ayant une orientation similaire.




(a) (b)

FIGURE 4.3 – Résultats de l'ensemble de données KTH. (a) Echantillon d'actions, (b) Matrice de confusion utilisant un bloc de $5 \times 5$

*Base vidéo ADL (Activities of Daily Living - Activités de la Vie Quotidienne) [MPK09] :* C'est une base de vidéos haute définition ($1280 \times 720$ pixels) regroupant 10 actions courantes du quotidien (ex : peelBanana, useSilverware, answerPhone) effectuées par 5 personnes différentes. Nous suivons le protocole "leave-one-out" dans notre expérimentation. Pour cela, nous prenons en compte une séquence en tant que séquence requête, et toutes les autres comme séquences de




(a) (b)

FIGURE 4.4 – Résultats pour la base ADL. (a) Echantillons d'actions. (b) Matrice de confusion utilisant des blocs de $5 \times 5$ pixels

référence pour la phase de reconnaissance d'action. Cette procédure est effectuée pour toutes les séquences, et la moyenne des résultats est calculée pour chaque classe d'action.

Dans la Figure 4.4, nous présentons la matrice de confusion obtenue dans le cadre de notre approche avec cette base de vidéos. L'approche obtient une précision moyenne de 0.84 pour des blocs de taille $5 \times 5$ pixels. Ce résultat est très encourageant, toutefois, l'action "peelBanana" peut être confondue avec les actions "eatSack" et "useSilverware" car elle a un comportement initial similaire qui consiste à ramener un objet depuis le potager.

### 4.4.2 Étude comparative

Nous comparons notre approche avec d'autres en utilisant les bases de vidéos KTH et ADL, et présentons leurs précisions dans le tableau 4.1.

Il s'avère que les approches basées sur les descripteurs spatiotemporels locaux [DRCB05, LMSR08] et l'historique des vitesses [MPK09] aboutissent à de meilleurs résultats que notre système pour la base KTH. Ce dernier a recours à la vitesse des points d'intérêt en tant que

| Méthode | KTH | ADL |
|---|---|---|
| Approche proposée | 0.58 | 0.84 |
| Historique des vitesses [MPK09] | 0.74 | 0.63 |
| Points d'intérêt spatiotemporels [LMSR08] | 0.80 | 0.59 |
| Cuboïdes spatiotemporels [DRCB05] | 0.66 | 0.36 |

TABLE 4.1 – Comparaison pour 2 bases de vidéos

descripteurs de bas niveau. Cependant, notre système est plus performant avec la base ADL car il combine à la fois les informations relatives à la magnitude du mouvement et à l'orientation.

Par rapport aux caractéristiques HOG/HOF exploitées par [LMSR08], notre modèle de scène assimile les principales orientations et magnitudes, et il ne prend pas en compte les mouvements soumis au bruit. De plus, chaque mélange de lois comprend des orientations moyennes avec les variances et poids correspondants, tandis que les descripteurs HOG/HOF calculent des histogrammes sur des gradients orientés (HOG) et des flux optiques (HOF) qui sont moins précis. Notre approche est notamment efficace lors de l'utilisation de la base de vidéos de haute résolution ADL car elle s'appuie sur l'information de mouvement qui est plus exacte. Néanmoins, elle souffre d'un manque de précision sur les vidéos basse résolution de la base KTH.

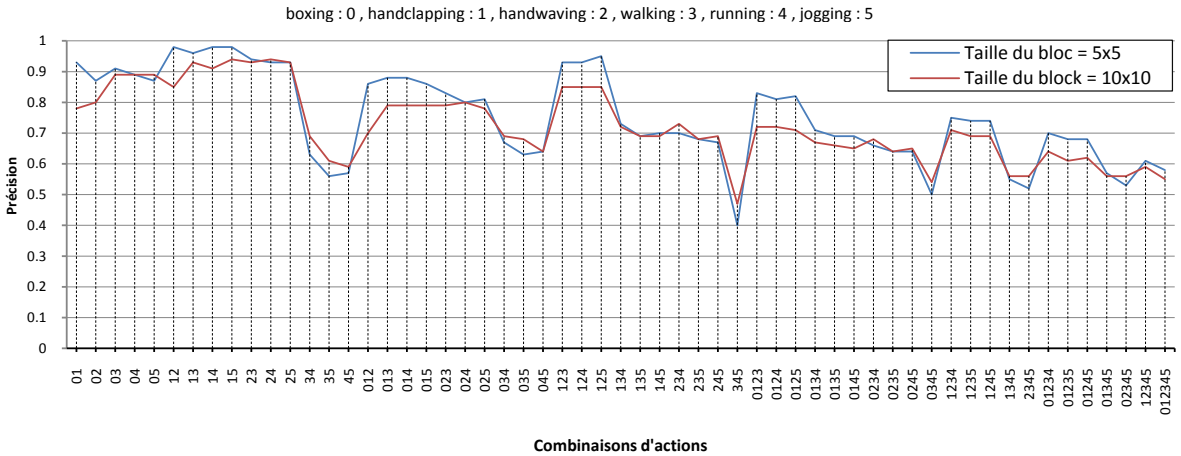



FIGURE 4.5 – Influence de la taille des blocs et de la combinaison des actions sur la précision

### 4.4.3 Étude du nombre de classes d'actions et de la taille des blocs

Nous étudions l'influence de la taille des blocs et le nombre de classes d'action sur la base vidéo KTH. Ainsi, nous avons répété l'expérience pour chaque élément de l'ensemble des sous-ensembles des actions de la base KTH pour les actions suivantes : handshaking, boxing, handwaving, walking, running et jogging. Nous avons respectivement noté ces actions $A = \{0,1,2,3,4,5\}$. Les graphes de la Figure 4.5 montrent la précision de notre système pour chaque sous-ensemble de $A$. Le graphe bleu est obtenu pour des blocs de taille $5 \times 5$ pixels, tandis que le graphe rouge correspond à des blocs de taille $10 \times 10$ pixels.

Le taux de précision le plus bas $(\sim 40\%)$ est atteint lorsque les actions 345 sont combinées (correspondant aux actions walking, running et jogging). Cela souligne la difficulté à différencier la vitesse de chaque action dans le cadre de vidéos basse résolution, car le mouvement relatif au niveau des pixels est très similaire à cause de la faible résolution.

Nos expériences montrent également que le fait d'augmenter la taille des blocs réduit la précision globale du système. Cependant, le temps de traitement est lui aussi diminué. Par ailleurs, l'augmentation du nombre de séquences de référence allonge la durée du traitement.

## 4.5 Conclusion

Nous avons présenté dans ce chapitre un système de reconnaissance d'actions performant qui se base sur les modèles de direction et les modèles de magnitude du mouvement. Notre approche a suivi une méthodologie en trois niveau qui permet le passage du signal vidéo à la reconnaissance de l'action exécutée.

Dans le premier niveau, nous avons extrait les vecteurs de flux optique des séquences vidéo. Ces vecteurs nous ont permis d'estimer des modèles statistiques sur l'orientation et la magnitude du mouvement constituant le niveau intermédiaire. Le résultat est un modèle de séquence vidéo qui estime les principales orientations et magnitudes dans tous les blocs de la scène. Le niveau sémantique de notre approche est atteint en utilisant une mesure de distance. Cette dernière

permet de comparer le modèle d'une séquence à des modèles de référence afin de reconnaitre l'action.

En s'appuyant sur l'orientation et la magnitude du mouvement, notre approche aboutit à des résultats prometteurs en comparaison avec d'autres approches de l'état de l'art, notamment sur des vidéos en haute définition.

# Chapitre 5

# Analyse du comportement humain dans une scène de foule

## 5.1 Extraction des motifs de mouvement

Les motifs de mouvement (*motion patterns*, en anglais) sont largement utilisés pour la modélisation des comportements habituels dans une scène. L'extraction de motifs de mouvement consiste à trouver les trajectoires ou les modalités de mouvement dominantes pour chaque région de la scène. Les motifs de mouvement sont généralement représentés par une image dessinant les différentes trajectoires par une couleur différente. La Figure 5.1 illustre trois motifs de mouvement extraits d'une vidéo du jeu de données Caviar [5]. Chaque image noire représente un motif de mouvement schématisé par une flèche et des blocs de couleurs différentes. La flèche représente la direction du mouvement principale du motif et la couleur d'un bloc représente l'orientation du mouvement principale des blocs au sein du motif. La correspondance couleur/orientation est effectuée à l'aide du guide situé dans le coin inférieur droit de la figure.

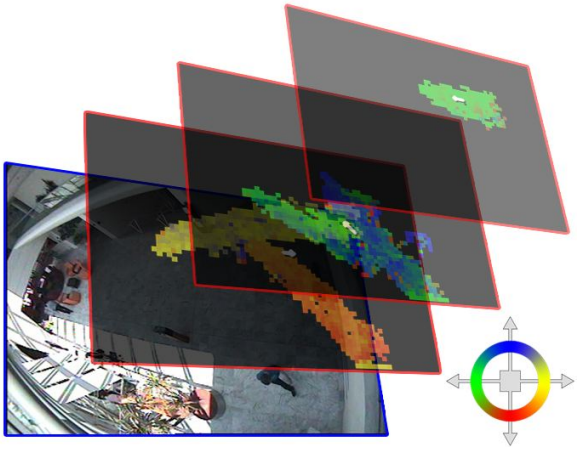


FIGURE 5.1 – Motifs de mouvement

La plupart des approches traditionnelles de détection des motifs de mouvement supposent que la scène soit structurée et ne prennent pas en compte les scènes complexes. Une scène est *structurée* quand les objets ne se déplacent pas de façon aléatoire et suivent une seule direction (ou modalité) de mouvement pour chaque région de la scène. A l'opposé, une scène *non structurée* peut contenir divers types de mouvements dans au moins un région de la scène.

5. http ://homepages.inf.ed.ac.uk/rbf/CAVIAR/

Notre approche traite aussi bien les scènes structurées que non structurées. Elle est décomposée sur trois niveaux : niveau bas (estimation du flux optique), niveau intermédiaire (construction du modèle directionnel) et niveau sémantique. Nous explicitons le dernier niveau dans la Section 5.1.1. Son but est de regrouper les cellules constituant le modèle directionnel où chaque groupe forme un motif de mouvement. La Section 5.1.2 décrit une méthode permettant d'exploiter le modèle directionnel pour extraire les motifs de mouvement. La Section 5.1.3 présente les résultats expérimentaux de notre approche.

### 5.1.1 Description de l'approche

Notre approche détecte les motifs de mouvement à partir d'une séquence vidéo par une démarche pyramidale en trois niveaux (introduite dans la Section 1.5). La Figure 5.2 montre le rôle de chaque niveau :

- Niveau bas : il a pour but d'estimer le mouvement dans la scène. Nous utilisons les points d'intérêt (voir Section 2.2.3) auxquels nous appliquons la méthode de flux optique décrite dans la Section 2.3.1. La méthode de calcul de flux optique choisie est bien adaptée pour les scènes de foule car elle est basée sur l'analyse du mouvement des points d'intérêt et ne dépend pas du nombre d'objets présents dans la scène. Une approche d'extraction de l'arrière-plan ne serait pas adaptée puisque les occlusions ne permettent pas de détecter et de suivre efficacement les personnes.
- Niveau intermédiaire : nous calculons le modèle directionnel présenté dans la Section 3.2 car il permet de représenter efficacement les orientations dominantes du mouvement dans chaque région de la scène. En effet, chaque cellule du modèle directionnel peut contenir plusieurs orientations dominantes. Cependant, les motifs de mouvement n'y sont pas clairement définis.
- Niveau sémantique : nous introduisons un nouvel algorithme qui permet de regrouper des zones voisines du modèle directionnel correspondant à des régions de la scène qui affichent des mouvements d'orientations similaires. Chaque groupe correspond alors à un motif de mouvement. L'avantage de cette approche est qu'une région de la scène peut

représenter des motifs de mouvement différents. Ceci est possible grâce au modèle directionnel car il extrait un ensemble d'orientations dominantes où chacune est susceptible d'appartenir à un motif de mouvement. Cet avantage permet notamment à notre approche d'extraire, dans les scènes non structurées, des motifs différents dans une même zone. Nous détaillons cette étape dans la section suivante.

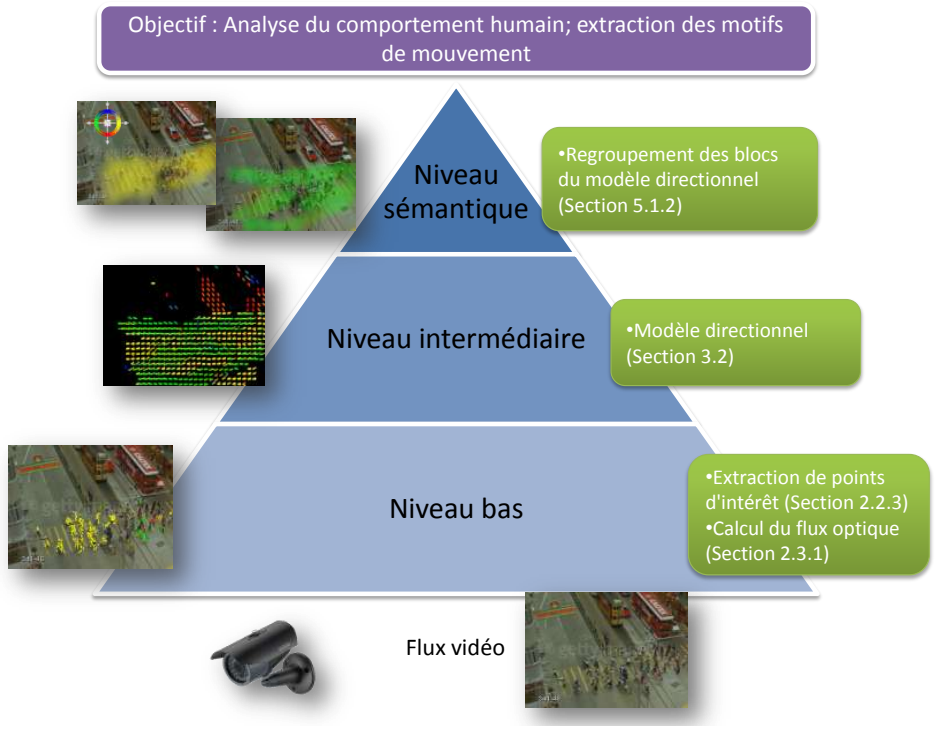



FIGURE 5.2 – Schéma de notre approche d'extraction des motifs de mouvement

### 5.1.2 Extraction des motifs de mouvement (niveau sémantique)

Cette étape a pour but de calculer les motifs de mouvement à partir du modèle directionnel d'une vidéo. Tout d'abord nous estimons un modèle directionnel qui contient $Bx \times By$ blocs ; où chaque bloc contient jusqu'à $K$ orientations majeures.

Le problème d'extraction de motifs de mouvement peut alors être formulé comme un problème de regroupement des blocs du modèle directionnel. Nous nous référons au gestaltisme [Ste03] afin de trouver les critères de regroupement tels que la proximité, la similitude, la bonne

continuité et le destin commun. Ensuite, nous détectons les motifs de mouvement de la scène en appliquant un algorithme de segmentation en régions sur les blocs du modèle directionnel. L'algorithme complet est présenté ci-dessous. La Figure 5.3 montre le résultat de notre algorithme sur un modèle directionnel ($3 \times 3$) avec $K = 4$.

Notre algorithme d'extraction de motifs de mouvement (voir Algorithme 3) regroupe deux blocs voisins s'ils ont au moins deux orientations similaires parmi les $K$ orientations estimées dans chacun des deux blocs. Par conséquent, au moins une des $K$ orientations principales du premier bloc doit être similaire a au moins une des $K$ orientations principales du deuxième bloc. Ainsi, un bloc peut appartenir à $K$ groupes au maximum. Cette possibilité est une particularité de notre algorithme comparé aux algorithmes de regroupement classiques, où un élément n'appartient qu'à un seul groupe à la fois. Ceci est réalisé en mémorisant le groupe correspondant à chacune des $K$ orientations dominantes. Pour cela, nous utilisons une matrice 3D (Bx $\times$ By $\times K$) où chaque élément contient l'identifiant de groupe associé à une orientation dominante du modèle directionnel.

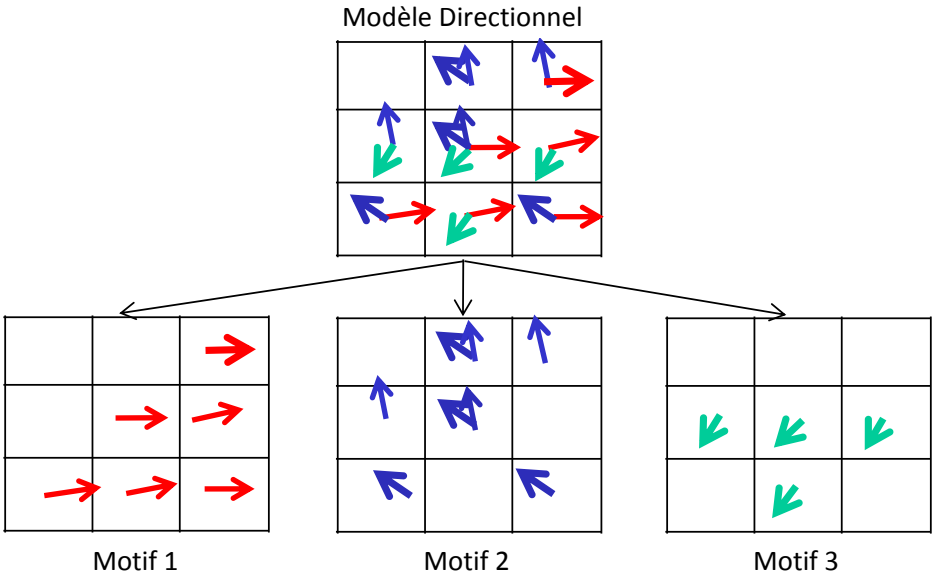



FIGURE 5.3 – Motifs de mouvement à partir d'un modèle directionnel de 3 lignes et 3 colonnes

**Algorithme 3** Algorithme d'extraction des motifs de mouvement depuis un modèle directionnel

**Entrée** Modèle directionnel D qui contient $Bx \times By$ mélanges de $K$ lois de von Mises
**Sortie** Ensemble de groupes qui représentent les motifs de mouvement
Créer une matrice 3D $M$ de dimensions $Bx \times By \times K$. $M(i,j,l)$ contient l'identifiant du groupe qui lui sera affecté
Créer une matrice 3D $\mu$ de dimensions $Bx \times By \times K$ et initialiser $\mu(i,j,l)$ avec la direction modale de la $l^{\text{ème}}$ loi de von Mises du bloc $(i,j)$
Initialiser l'ensemble des groupes : $C \leftarrow \emptyset$
Initialiser la variable qui indique le groupe courant : $n \leftarrow 0$
Initialiser la matrice 3D d'identifiants : $M \leftarrow 0$
**pour** $i = 1$ à $Bx$ **faire**
  **pour** $j = 1$ à $By$ **faire**
    **pour** $l = 1$ à $K$ **faire**
      **si** $M(i,j,l) = 0$ **alors**
        $n \leftarrow n+1$
        créer un nouveau groupe c
        insérer l'élément $(i,j,l)$ de direction modale $\mu_{i,j,l}$ dans c
        $C \leftarrow C \cup c$
        $B \leftarrow voisins(i,j,l,M)$
        $M(i,j,l) = n$
        **pour** chaque $b$ dans $B$ **faire**
          **si** $directionMoyenne(c) - \mu(b.x, b.y, b.k) \leq \alpha$ **alors**
            $M(i,j,l) = n$
            insérer l'élément $(b.i, b.j, b.l)$ de direction modale $\mu_{b.x,b.y,b.k}$ dans c
            $B \leftarrow B \cup voisins(b.x, b.y, b.k, M)$
          **fin si**
        **fin pour**
      **fin si**
    **fin pour**
  **fin pour**
**fin pour**

### 5.1.3 Expérimentations

Notre approche d'extraction des motifs de mouvement a été expérimentée sur des scènes structurées et non structurées. Les scènes structurées sont simples à traiter car les objets présents se déplacent tous de la même manière. Les scènes non structurées sont caractérisées par un certain désordre (ou chaos) où les objets ne se déplacent pas de façon organisée sur la scène. Pour extraire les motifs de mouvement sur ces scènes, nous estimons les vecteurs de flux optique,

afin de construire un modèle directionnel. L'algorithme d'extraction de motifs de mouvement est alors exécuté sur ce modèle directionnel.

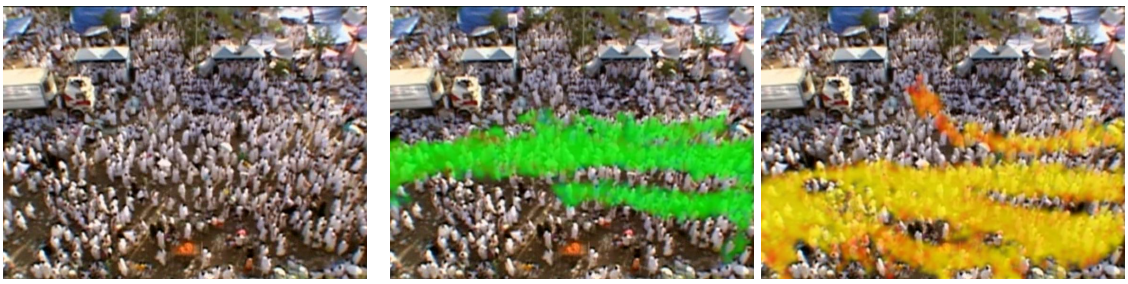
Notre approche a d'abord été évaluée dans un environnement urbain où les véhicules et les piétons utilisent la même route, comme illustré dans la Figure 5.4(a). La séquence vidéo utilisée provient de la base de données AVSS 2007 [6]. Elle a une résolution de $720 \times 576$ pixels avec un taux d'images de 25 images/seconde. La scène est non structurée. E lle représente une route à deux voies, les voitures circulent à gauche sur la chaussée. Des véhicules empruntent la route et quelques piétons la traversent sur le passage piéton en bas de la scène, matérialisé par les deux bornes claires. Notre approche parvient à extraire les motifs de mouvement des véhicules en distinguant deux motifs correspondant au trafic sur la route principale et un troisième motif pour les voitures qui tournent à gauche. Elle a également mis en évidence les motifs de mouvement des piétons en bas de la scène. La Figure 5.4(b) illustre chaque motif de mouvement dans une image différente. Chaque couleur a une orientation donnée par le cercle coloré en haut à gauche de la Figure 5.4(a). La capacité d'affecter plusieurs groupes à un seul bloc permet d'identifier le motif généré par les piétons. Les approches qui se limitent à une orientation unique pour chaque région dans la scène négligent le motif généré pas les piétons.


(a) (b)

FIGURE 5.4 – Motifs de mouvement détectés dans une scène urbaine. (a) Échantillon de la séquence, (b) Motifs de mouvement extraits. Ils sont mieux visibles sur un document en couleurs

6. http ://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_d.html

La Figure 5.5 montre une autre scène non structurée. Dans cette vidéo, une importante quantité de personnes parcourt la scène dans différentes directions. Cependant, l'approche permet d'extraire deux motifs de mouvements, malgré le mouvement de chaos apparent. La recherche en intelligence collective vient consolider cette constatation car elle stipule que les organismes qui se déplacent d'une manière qui apparaît aléatoire ou chaotique tendent à se déplacer de façon organisée à travers le temps.



(a) (b)

FIGURE 5.5 – Motifs de mouvements détectés dans une scène de pèlerinage. (a)Échantillon de la séquence, (b) Motifs de mouvements extraits

Nous comparons notre approche avec celle de Hu et al. [HAS08b] par rapport à la séquence précédente. Ces derniers ont proposé une méthode d'extraction de motifs de mouvement en regroupant l'ensemble des vecteurs de flux optique de la séquence (aussi appelé champ de mouvement). Nous montrons les résultats de cette l'approche dans la Figure 5.6. Nous remarquons que notre approche a de meilleurs résultats car elle différencie les mouvements qui se chevauchent, contrairement à l'approche de [HAS08b] où les motifs en marron et orange ne se chevauchent pas, notamment dans la partie droite de la scène. On remarque aussi que l'approche de Hu et al. détecte moins de mouvements dans la partie supérieure de la scène, car elle utilise une étape de prétraitement qui peut éliminer l'information de mouvement utile et ainsi fausser l'extraction des motifs de mouvement. Cette étape de prétraitement est nécessaire en raison du grand nombre de vecteurs dans le champ de mouvement, contrairement à notre approche qui élimine les informations moins fréquentes à la volée, et ce, lors de la construction du modèle directionnel.

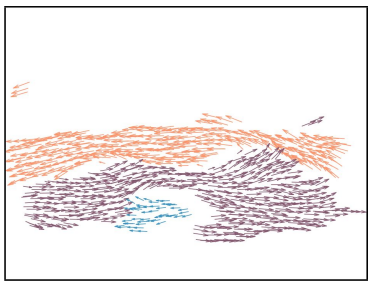


FIGURE 5.6 – Les motifs de mouvement détectés par l'approche de Hu et al. [HAS08b] sur la scène de pèlerinage. Chaque couleur indique un motif différent

(a) Échantillon de la séquence

(b) Motifs de mouvement

FIGURE 5.7 – Les motifs de mouvement extraits d'une scène routière complexe

Nous illustrons d'autres résultats de notre approche dans les Figures 5.8 et 5.9. Elles proviennent des bases de vidéos CAVIAR [7] et Getty-images [8] caractérisées par des foules denses ayant un comportement non structuré. Nous pouvons remarquer l'importance des motifs de mouvement en termes d'analyse du comportement car elles permettent de synthétiser le comportement des foules durant une période de temps.

Finalement, nous synthétisons les résultats de nos expérimentations dans le tableau 5.1. Il compare le nombre de motifs de mouvement détectés avec la vérité terrain. Les noms des

7. http ://homepages.inf.ed.ac.uk/rbf/CAVIAR/
8. http ://www.gettyimages.com/

(a) Échantillon de la séquence (b) Motifs de mouvements

FIGURE 5.8 – Les motifs de mouvement extraits d'une autre séquence de pèlerinage

(a) Échantillon de la séquence

(b) Motifs de mouvement

FIGURE 5.9 – Les motifs de mouvement extraits à l'entrée d'un escalator

| Video | Motifs de mouvement détectés | Vérité terrain |
|---|---|---|
| AVSS_PV_Hard | 4 | 4 |
| pilgrimage1 | 2 | 2 |
| pilgrimage2 | 3 | 2 |
| Caviar_Browse4 | 4 | 4 |
| Caviar_EnterExitCrossingPath | 3 | 3 |
| Getty-341-46_l | 3 | 2 |
| Getty-81059009_q | 5 | 5 |

TABLE 5.1 – Comparaison des résultats de notre approche avec la vérité terrain

séquences vidéo sont ceux des fichiers d'origine. Malgré la complexité des séquences, notre approche fournit des résultats très pertinents y compris au niveau des zones de chevauchement de plusieurs motifs de mouvement.

## 5.2 Détection d'évènements de foule

La détection d'évènements se définit par la détection de situations attirant l'attention humaine [SaMCC08]. C'est un domaine très large en raison du grand nombre d'évènements possibles et des différents scénarios pouvant aboutir à leur réalisation. Par ailleurs, la définition-même du terme "évènement" dans la vidéo varie selon le contexte. De nombreux efforts ont été entrepris dans le domaine de la détection d'évènements *anormaux* ou *inhabituels* dans des scènes de foule. Cependant, il existe peu de travaux consacrés à la classification de plusieurs types d'évènements dans de telles scènes.

Dans ce qui suit, nous utilisons le modèle directionnel pour représenter l'environnement. Le modèle permet de gérer de manière effective la diversité et la complexité des scénarios. Ces modèles sont utilisés pour extraire un ensemble de caractéristiques liées au comportement de la foule. Ces caractéristiques servent à entrainer un classificateur pour déduire quel type d'évènement s'est produit. Notre méthode a été appliquée à une sélection d'évènements et a été évaluée sur la base *PETS'2009* [9].

### 5.2.1 Description de l'approche

Notre méthode suit une démarche pyramidale en trois niveaux illustrée dans la Figure 5.10.

- Niveau bas : La première étape porte sur l'extraction d'une série de points d'intérêt tel que défini dans la Section 2.2.3. Ces points sont par la suite suivis grâce à des techniques de calcul du flux optique. Nous avons choisi la méthode mentionnée dans la Section 2.3.1 Les points statiques sont éliminés afin de se concentrer sur les zones en mouvement.
- Niveau intermédiaire : Le modèle directionnel caractérisant l'orientation du mouvement dans chaque bloc est estimé pour chaque image. Les blocs voisins sont ensuite regroupés selon le degré de similarité des orientations afin de segmenter des groupes de personnes se déplaçant dans la même direction. Ces groupes sont ensuite suivis dans les prochaines images afin de détecter leur disparition ou leur séparation. Voir les Sections 3.2 et 3.3.2

9. http ://www.cvg.rdg.ac.uk/PETS2009/

pour plus de détails sur l'estimation du modèle directionnel et la méthode de détection et de suivi des groupes de personnes.

– Niveau sémantique : Les évènements sont détectés grâce aux informations obtenues lors du suivi des groupes. Notre approche permet de définir des seuils expérimentalement pouvant être édités à la volée ou d'utiliser des classificateurs pour détecter les évènements.

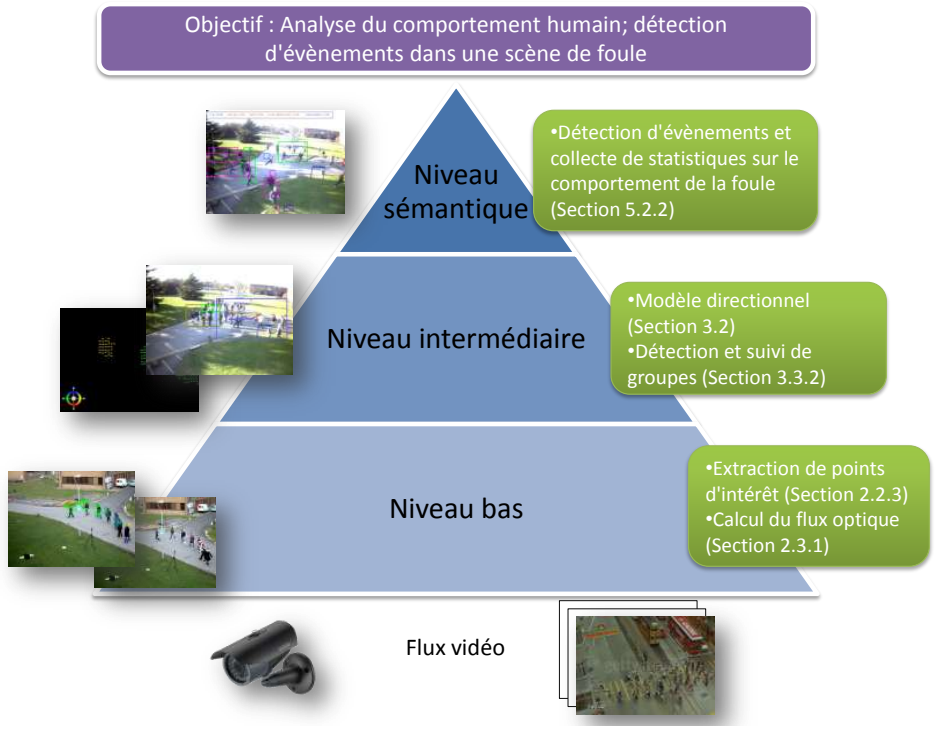



FIGURE 5.10 – Schéma de notre approche de détection d'évènements de foule

### 5.2.2 Détection d'évènements de foule (niveau sémantique)

Dans cette section, nous décrivons la détection d'évènements de foule. Les scénarii choisis font partie des évènements décrits dans la base PETS'2009. Ils sont au nombre de six que nous divisons en deux catégories :

– Course et marche : il s'agit de déterminer si les personnes formant un groupe sont en train de marcher ou de courir. Ces évènements peuvent être identifiés en utilisant la magnitude des vecteurs du flux optique.

– Fusion et division : Il s'agit de déterminer si les groupes de personnes se rejoignent ou se séparent. Le workshop PETS'2009 définit trois types de séparation :

1. La dispersion locale ; où deux ou plusieurs groupes de la foule émergent de la foule en se déplaçant dans des directions différentes sans pour autant se séparer complètement.
2. La séparation ; où les groupes ayant entamé une dispersion locale se séparent réellement de la foule et empruntent des directions de mouvement différentes.
3. L'évacuation est un cas particulier de la séparation où la foule se divise subitement en courant.

Les évènements appartenant à la première catégorie sont détectés en comparant la magnitude moyenne du flux optique de chaque image avec la vitesse moyenne de la scène préalablement obtenue par apprentissage. Les évènements de la seconde catégorie sont détectés en analysant la position, l'orientation et la vitesse des groupes. Une explication plus détaillée est donnée dans les sections suivantes.

#### Course et marche

L'idée principale consiste à calculer la magnitude moyenne des vecteurs de mouvement dans chaque image. Une magnitude élevée signifie l'évènement course tandis qu'une magnitude faible signifie l'évènement marche. La détection se fait soit en définissant un seuil expérimental ou bien en utilisant un classificateur avec comme caractéristique la vitesse moyenne de mouvement.

#### Fusion et division

Pour détecter les évènements *fusion* et *division*, nous calculons la variance circulaire $S_{0,f}$ relative aux orientations de déplacement des groupes dans chaque image $f$ selon l'équation suivante [GB80] :

$$S_{0,f} = 1 - \frac{1}{n_f} \sum_{i=1}^{n_f} \cos\left(X_{i,f} - \overline{X_{0,f}}\right) \tag{5.1}$$

où $\overline{X_{0,f}}$ est l'angle moyen des groupes dans l'image $f$ définie par :

$$\overline{X_{0,f}} = \arctan \frac{\sum_{i=1}^{n_f} \sin(X_{i,f})}{\sum_{i=1}^{n_f} \cos(X_{i,f})} \tag{5.2}$$

La variance circulaire $S_{0,f}$ représente la dispersion des groupes. Elle est comprise entre 0 et 1 inclus. $S_{0,f}$ vaut 0 pour un ensemble d'angles identiques et vaut 1 pour un ensemble d'angles totalement opposés. On peut alors inférer un évènement de fusion ou de division d'après la valeur de $S_{0,f}$ en utilisant un seuil expérimental $S_{0,seuil}$ ou un classificateur.

L'étape suivante consiste à déterminer quel évènement de fusion ou de division s'est produit, car la variance circulaire ne suffit pas pour les différencier. Nous examinons pour cela la position et l'orientation de chaque groupe par rapport aux autres groupes. Si deux groupes sont orientés vers la même direction et sont proches les uns des autres, alors il s'agit d'une fusion. Toutefois, s'ils se dirigent dans des directions opposées tout en étant proches les uns des autres, alors il s'agit d'une dispersion.

Soit $\overrightarrow{v_{i,f}}$ un vecteur représentant le mouvement du groupe $C_{i,f}$ à l'image $f$. Le vecteur $\overrightarrow{v_{i,f}}$ est défini par son origine $O_{i,f}$ qui est le centre du groupe $C_{i,f}$, une direction $\Omega_i$ et son point d'arrivée $Q_{i,f}$ de coordonnées $qx_{i,f}, qy_{i,f}$ définies par les équations suivantes :

$$qx_{i,f} = ox_{i,f} \cdot \cos(\Omega_i) \tag{5.3}$$

$$qy_{i,f} = oy_{i,f} \cdot \sin(\Omega_i) \tag{5.4}$$

On peut désormais détecter un évènement de fusion ou de division entre deux groupes $i$ et $j$ grâce aux distances entre les points $O_i$ et $O_j$ et les points $Q_i$ et $Q_j$. Nous avons choisi d'utiliser la distance euclidienne pour sa simplicité et sa rapidité pour les applications temps réel. Comme

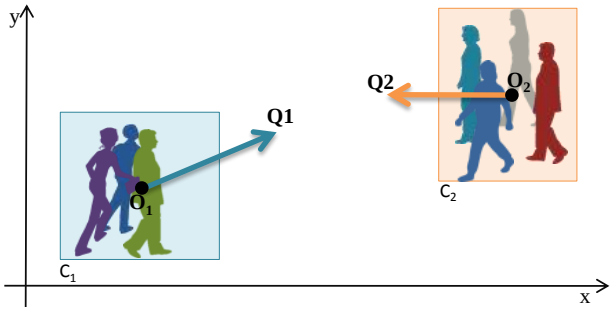



FIGURE 5.11 – Représentation de groupes en fusion

pour les cas précédents, nous utilisons ces métriques dans un classificateur et nous proposons aussi des seuils expérimentaux comme suit. Deux groupes $i$ et $j$ fusionnent si :

$$\begin{cases} D(O_i, O_j) > D(Q_i, Q_j) \\ et \\ D(O_i, O_j) < \delta \end{cases} \tag{5.5}$$

où $D(P,Q)$ est la distance euclidienne entre les points $P$ et $Q$ et $\delta$ est un seuil fixé expérimentalement et représente la distance minimale entre deux groupes. La Figure 5.11 illustre deux groupes en fusion.

De façon similaire, deux groupes se divisent si les conditions suivantes sont satisfaites :

$$\begin{cases} D(O_i, O_j) < D(Q_i, Q_j) \\ et \\ D(O_i, O_j) < \delta \end{cases} \tag{5.6}$$

Comme explicité précédemment, il y a trois situations distinctes dans le cas d'une *division*. Ces situations illustrées dans les Figures 5.12 et 5.13 sont :

1. La division se produit dans une zone restreinte pendant une courte durée. Il s'agit alors de *dispersion locale*.
2. La division se produit dans une zone plus grande avec une divergence considérable des groupes. Il s'agit alors d'une *séparation*.

3. Si la première situation se produit alors que la foule est en train de courir, il s'agit d'une *évacuation*.

Afin de distinguer ces évènements, nous calculons un 'age' pour chaque groupe $C_{i,f}$ qui n'est autre que le numéro de l'image où le groupe a été détecté pour la première fois, que nous notons $F_{i,f}$. Il y a une dispersion locale à l'image $f$ entre deux groupes $C_{i,f}$ et $C_{j,f}$ si les condition de l'équation 5.6 sont satisfaites et si leur age est petit :

$$\begin{cases} f - F_{i,f} < \nu \\ and \\ f - F_{j,f} < \nu \end{cases} \tag{5.7}$$

où $\nu$ est un seuil défini expérimentalement.

Deux groupes $C_{i,f}$ et $C_{j,f}$ se séparent à l'image $f$, si les conditions de l'équation 5.6 sont satisfaites. D'autre part, l'un des deux groupes doit avoir un âge plus grand :

$$\begin{cases} f - F_{i,f} \geq \nu \\ or \\ f - F_{j,f} \geq \nu \end{cases} \tag{5.8}$$

L'évolution de la division d'un groupe à la séparation en passant par une dispersion locale est représentée dans la Figure 5.12.

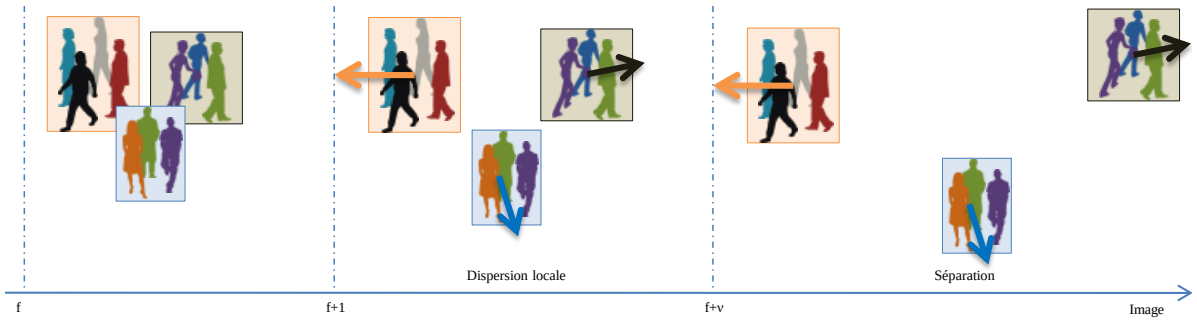



FIGURE 5.12 – Représentation des évènements dispersion locale (au milieu) et séparation (à droite)

Un évènement d'évacuation correspond à une dispersion locale affectant deux groupes caractérisés simultanément par un évènement de course comme l'illustre la Figure 5.13.

FIGURE 5.13 – Représentation de l'évènement évacuation

En résumé, notre approche définit un ensemble de descripteurs qui modélisent le comportement de la foule. Nous proposons deux façons de mettre en œuvre la détection d'évènements à partir de ces descripteurs ; Soit à l'aide de seuils manuels, soit à l'aide de classificateurs. Dans le cas de l'utilisation de classificateurs, nous gardons les descripteurs suivants : la vitesse moyenne de la foule, le nombre de groupes, leur direction moyenne, leur variance circulaire, les distances moyennes $D(O_i, O_j)$ et $D(Q_i, Q_j)$ entre chaque couple de groupes, ainsi que la moyenne et la variance de l'âge des groupes.

### 5.2.3 Expériementations et résultats

Cette section décrit les résultats de l'expérimentation ainsi qu'une comparaison de nos résultats avec deux autres approches.

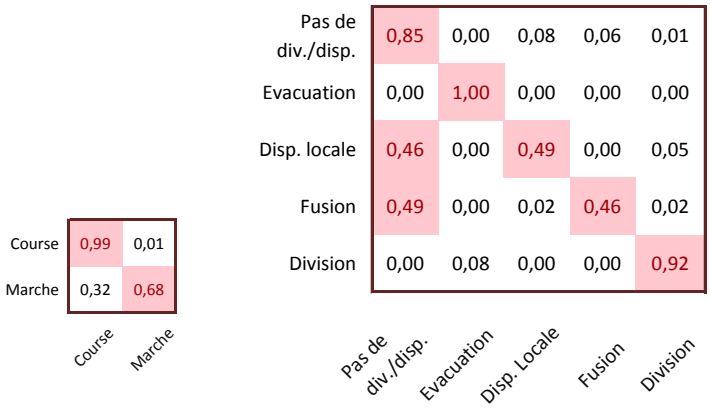
#### Expérimentation

L'approche décrite dans les sections précédentes a été évaluée sur la base de vidéos PETS'2009. Ces vidéos incluent des séquences contenant différentes situations de foule. Nous y trouvons des scénarios impliquant le calcul de la densité de foule, le comptage du nombre de personnes, le suivi des personnes, l'analyse des motifs de mouvement et la détection d'évè-

nements sémantiques. Ces séquences sont au format JPEG avec une résolution de $720 \times 576$ pixels et une cadence d'environs 8 images par seconde.

Pour la classification, nous utilisons deux classificateurs car chaque image a deux classes, une pour chaque catégorie d'évènements (course ou marche, division ou fusion). Bien que cela rende le système plus compliqué, nous avons néanmoins la possibilité de distinguer les évènements liés à la vitesse et ceux liés à la division et à la fusion. Nous avons divisé la base de vidéos en deux ensembles, le premier est utilisé pour l'apprentissage et contient 75% du nombre total d'images (qui est approximativement 1000). Le deuxième est utilisé pour tester les deux classificateurs et contient les 25% d'images restants.

Parmi les classificateurs testés, le classificateur 'Forêt aléatoire' [CLB04] a donné les meilleurs résultats illustrés à travers les matrices de confusion dans la Figure 5.14. Celle-ci souligne une forte séparation entre les évènements et la faible proportion de fausses alertes. Cependant, des évènements n'ont pas été détectés en raison d'erreurs lors des étapes de bas niveau (comme l'acquisition d'image, la détection et le suivi des points d'intérêt etc.). Ces erreurs ont été provoquées par les zones d'ombre et le bruit.




(a) (b)

FIGURE 5.14 – Les matrices de confusion obtenues en utilisant le classificateur 'Forêt aléatoire', (a) les évènements marche et course, (b) Les évènements fusion, division, dispersion locale et évacuation

**Comparaison**

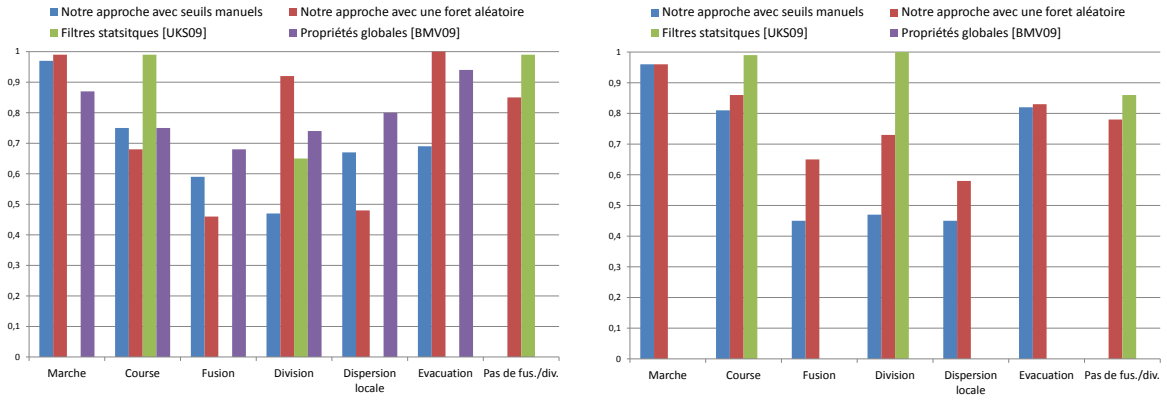
Nous regroupons les résultats obtenus en utilisant des seuils expérimentaux et des classificateurs ainsi que les résultats d'autres approches dans le Tableau 5.2. Il montre la précision et le rappel, le cas échéant, pour chaque évènement. En effet, certaines approches ne détectent pas certains évènements tandis que d'autres négligent de fournir certains résultats notés par NP sur le tableau. L'évènement 'régulier' a été rajouté et indique le cas où aucun évènement des deux catégories ne s'est produit. Une représentation graphique de ces résultats est illustrée dans la Figure 5.15 afin de faciliter leur lecture.

Nous notons que notre approche (avec seuils manuels ou classificateurs) est la seule qui soit capable de détecter tous les évènements. L'approche utilisant les filtres statistiques [UKS09] ne détecte que trois évènements, et l'approche qui étudie les propriétés globales des images [BMV09] ne considère pas l'évènement régulier en le confondant avec l'évènement marche.

L'approche d'Utasi et al. qui utilise les filtres statistiques [UKS09] a été conçue pour détecter des évènements 'anormaux' en utilisant les flux inhabituels de vitesse et de flot de mouvement comme descripteurs intermédiaires. Ces descripteurs ne peuvent détecter que trois catégories d'évènements (régulier, séparation et course). Cependant, les auteurs prétendent que leur approche est capable de détecter d'autres évènements anormaux en y greffant d'autres descripteurs. Les auteurs n'ont pourtant pas fourni plus de détails sur les démarches à suivre pour greffer d'autres descripteurs. En plus, nous pensons que des descripteurs de niveau intermédiaire conçus pour traiter les évènements sémantiques sont plus avantageux que ceux conçus uniquement pour des évènements anormaux. La table 5.2 montre que l'utilisation de classificateurs avec notre approche donne des résultats similaires à l'état de l'art avec l'avantage de traiter une grande palette d'évènements sans y ajouter de greffons.

Les résultats de notre approche sont très proches de l'approche utilisant les propriétés globales [BMV09]. Cependant, cette approche ne permet pas le chevauchement des évènements, ce qui signifie qu'elle ne peut pas détecter les évènements marche et fusion au même instant. Cette approche a aussi l'inconvénient d'être très gourmande en puissance de calcul car elle nécessite l'estimation d'une série temporelle pour chaque bloc d'images et ce, à chaque instant de

la vidéo. Notre approche quant à elle, atteint un débit de 4 images par seconde sur un processeur d'ordinateur portable de faible puissance de marque Intel Celeron 1.8 GhZ.




(a) Précision (b) Rappel

FIGURE 5.15 – Comparaions de différentes méthodes de détection d'évènements sous forme de graphique, (a) Précision, (b) Rappel

| | | Seuils manuels | Forêt aléatoire | Filtres statistiques [UKS09] | Propriétés globales [BMV09] |
|---|---|---|---|---|---|
| Marche | Précision | 0.97 | 0.99 | ND | 0.87 |
| | Rappel | 0.96 | 0.96 | ND | NC |
| Course | Précision | 0.75 | 0.68 | 0.99 | 0.75 |
| | Rappel | 0.81 | 0.86 | 0.99 | NC |
| Fusion | Précision | 0.59 | 0.46 | ND | 0.68 |
| | Rappel | 0.45 | 0.65 | ND | NC |
| Division | Précision | 0.47 | 0.92 | 0.65 | 0.74 |
| | Rappel | 0.47 | 0.73 | 1 | NC |
| Disp. locale | Précision | 0.67 | 0.48 | ND | 0.8 |
| | Rappel | 0.45 | 0.58 | ND | NC |
| Évacuation | Précision | 0.69 | 1 | ND | 0.94 |
| | Rappel | 0.82 | 0.83 | ND | NC |
| Pas de fus./div. | Précision | ND | 0.85 | 0.99 | ND |
| | Rappel | ND | 0.78 | 0.86 | ND |

TABLE 5.2 – Comparaions de différentes méthodes de détection d'évènements, NC signifie que les résultats n'ont pas été communiqués, ND signifie que l'évènement n'est pas détecté par l'approche


(a) Marche (b) Course

(c) Rassemblement (d) Séparation

(e) Dispersion locale (f) Evacuation

FIGURE 5.16 – Échantillon de détection d'évènements. Les évènements détectés apparaissent en bleu

## 5.3 Estimation des flux

L'estimation des flux consiste à compter le nombre de personnes traversant une zone prédéfinie. Le comptage de personnes a des applications intéressantes dans divers domaines allant du markéting à la vidéo-surveillance. Lorsque de nombreuses personnes traversent le hall d'une gare, d'un bâtiment ou d'un centre commercial, leur nombre représente une information pertinente pour les décideurs chargés d'assurer la sécurité publique et d'élaborer des stratégies commerciales. A titre d'exemple, le nombre de personnes qui entrent dans une salle peut dépasser une capacité qui est susceptible de refléter une situation anormale et potentiellement dangereuse. De plus, la possibilité de connaitre le nombre d'individus dans une zone commerciale représente une précieuse information pour les gérants d'un magasin afin d'évaluer l'attractivité des offres promotionnelles qu'ils proposent.

Les systèmes de comptage ont eu recours à des équipements spéciaux comme les barres de rotation ou les capteurs infrarouges. Ces derniers étaient mis en défaut par leur manque de précision notamment dans le cas où plusieurs personnes passent à la fois mais sont comptabilisées comme une seule personne. Un autre défaut notoire est qu'ils comptent également certaines personnes telles que les vigiles alors qu'il est préférable de les ignorer.

Par conséquent, les approches basées sur la vision par ordinateur sont plus appropriées pour élaborer des compteurs précis et bidirectionnels (qui détectent également la direction des personnes) car elles utilisent l'image qui permet d'interpréter de façon plus pertinente des situations complexes et d'ignorer certaines personnes (comme les vigiles qui sont habillés en rouge). Cependant, la plupart de ces approches ont été conçues pour une configuration d'environnement et de caméra précise et sont inefficaces en cas de changement de configuration.

Dans ce qui suit, nous présenterons une méthodologie d'estimation des flux indépendante de l'angle de prise de vue, permettant de compter le nombre d'individus franchissant une ligne de comptage à partir de vidéos issues de caméras monoculaires.

### 5.3.1 Description de l'approche

La plupart des systèmes de vidéo-surveillance utilisent des caméras installées d'une façon oblique pour obtenir une vue globale de la scène ainsi que certains détails tels que les visages et les vêtements. Cependant, afin d'éviter les occlusions et de respecter l'anonymat des individus, certaines caméras sont installées verticalement au-dessus de la tête des passants. La Figure 5.17 montre deux configurations de caméra pour lesquelles notre approche est adaptée.

(a) (b)

FIGURE 5.17 – Exemples de configuration de la caméra : (a) Caméra orientée verticalement au-dessus de la tête des passants, (b) Caméra orientée obliquement

Notre approche se divise en trois niveaux comme indiqué sur la Figure 5.18 à travers un schéma pyramidal. La Figure 5.19 illustre le résultat de chaque niveau.

- **Niveau bas :** cette étape a pour but la quantification du mouvement survenant sur une ligne de comptage préalablement définie (voir Section 2.2.1). Nous utilisons la méthode de calcul du flux optique présentée dans la Section 2.3.1. Le résultat de cette étape est constitué de deux cartes spatiotemporelles dont une estimée à partir des orientations et une autre estimée à partir des magnitudes des vecteurs de mouvement.
- **Niveau intermédiaire :** cette étape a pour but d'extraire les blobs correspondant aux personnes franchissant la ligne de comptage à partir des cartes spatiotemporelles détectées lors de l'étape précédente. La méthode de détection des blobs est basée sur le regroupe-

ment des pixels voisins sur la ligne de comptage ayant une orientation similaire. Elle a été décrite plus en détail dans la Section 3.3.1.

– **Niveau sémantique :** cette étape a pour but d'estimer le nombre de personnes dans un blob ainsi que sa direction. Les blobs détectés dans le niveau intermédiaire peuvent contenir une ou plusieurs personnes. Nous utilisons une méthode de régression linéaire pour déduire le nombre de personnes dans un blob à partir des ses caractéristiques. Nous décrivons cette étape dans la section suivante.

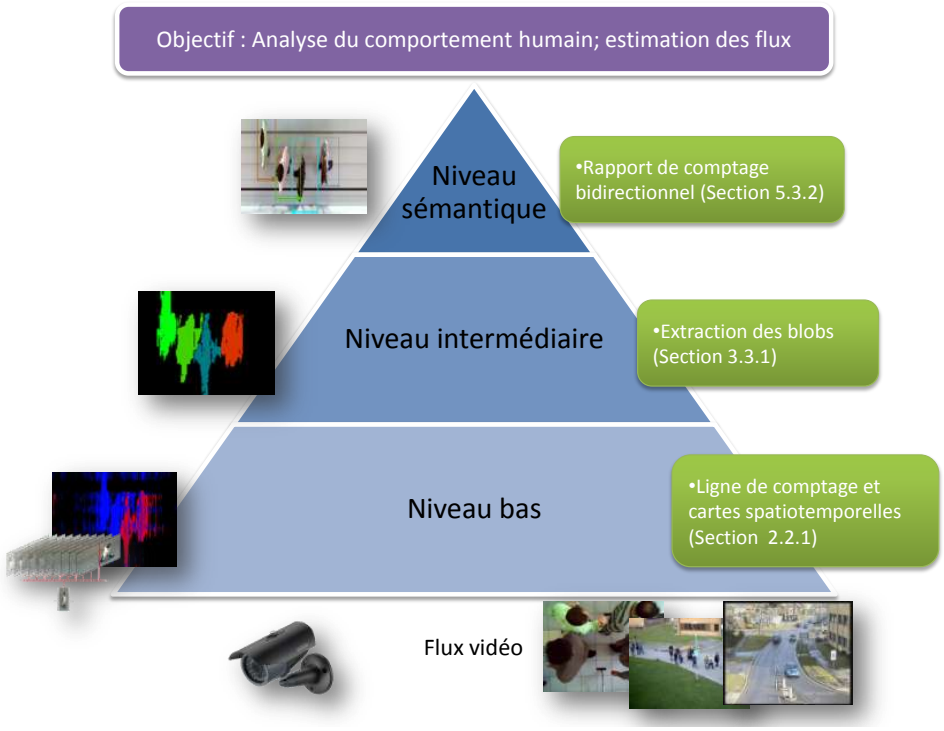
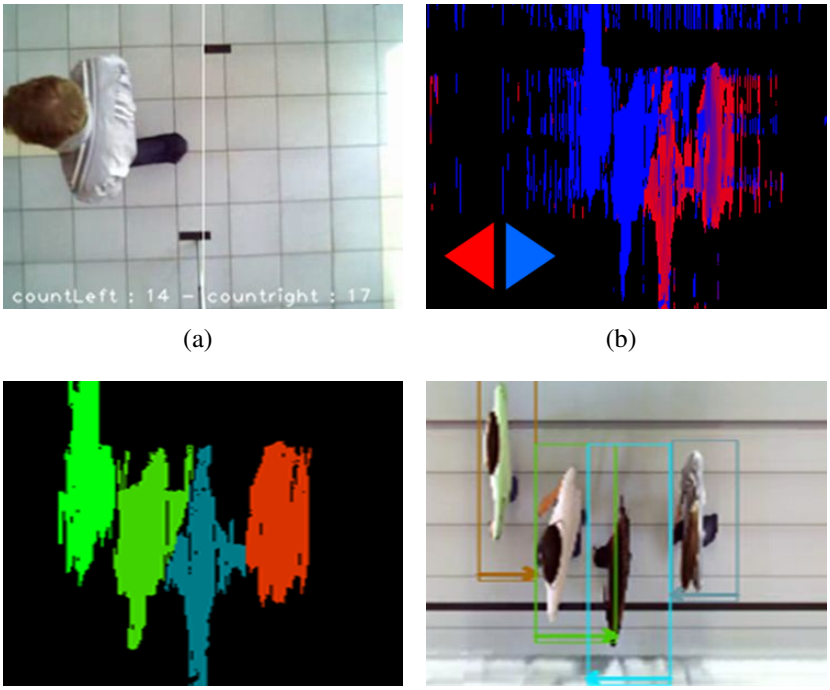



FIGURE 5.18 – Schéma de notre approche d'estimation des flux

### 5.3.2 Décision de comptage (niveau sémantique)

Les blobs détectés dans l'étape intermédiaire ne contiennent pas encore une information exacte de comptage, car ils peuvent contenir une ou plusieurs personnes. Le niveau sémantique a pour but d'estimer le nombre de personnes dans un blob et de restituer ainsi à l'utilisateur la direction et le nombre de personnes dans chaque blob. Pour cela, nous utilisons un modèle de




(a) (b)

(c) (d)

FIGURE 5.19 – Étapes clés de notre approche : (a) Sélection de la ligne de comptage, (b) Carte spatiotemporelle des orientations, (c) Détection en ligne des blobs, (d) Représentation de l'orientation des blobs sur la carte spatiotemporelle

régression linéaire pour estimer le nombre d'individus dans un blob. Le modèle est défini par la formule suivante :

$$y = H(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x \tag{5.9}$$

où :

– $y$ correspond à l'estimation du nombre d'individus dans un blob.
– L'hypothèse $H(x)$ correspond au modèle de régression linéaire.
– $n$ représente le nombre de caractéristiques d'un blob tel qu'indiqué dans la Formule 3.6 et qui sont $(P,N,\alpha,\beta,l,h,O,V)$. Pour rappel : $P$ est l'ensemble de pixels du blob et $N$ leur nombre. $l$ et $h$ sont respectivement la largeur et la hauteur du rectangle minimum englobant les pixels du blob. $\alpha$ et $\beta$ sont les coordonnées du coin supérieur gauche de ce rectangle. $O$ correspond à l'orientation moyenne du blob et $V$ à sa vitesse moyenne.

– $x$ est un vecteur de $n$ dimensions contenant les caractéristiques du blob.

– Le vecteur de $n$ dimensions des coefficients $\theta$ est estimé par un apprentissage supervisé.

Nous avons adopté une méthode d'apprentissage hors ligne pour chaque configuration différente en utilisant une vérité terrain annotée manuellement par un expert. Cela consiste en un ensemble de caractéristiques de blobs extraits automatiquement qui sont mis en correspondance avec le comptage d'individus effectué manuellement. L'utilisation d'un vecteur de coefficients distinct pour chaque configuration permet d'en appréhender d'autres facilement, car il suffit d'apprendre, pour une nouvelle configuration, le nouveau vecteur de paramètres pour que notre approche soit adaptée à la nouvelle configuration.

La méthode de régression *pace regression* est utilisée car elle a l'avantage, par rapport aux autres méthodes classiques de régression linéaire, d'évaluer l'effet de chaque variable et d'inclure cette information lors de l'évaluation du modèle de régression. De plus amples détails pour cette méthode sont disponibles dans [Wan00]. Cette étape d'apprentissage aboutit à un vecteur de coefficients $\theta$ correspondant à une configuration spécifique. Ce vecteur est introduit dans la Formule 5.9 qui renvoie une estimation de comptage en fonction du vecteur de caractéristiques $x$ d'un blob.

La direction d'un blob ("gauche" ou "droite" par exemple) dépend de l'orientation de la ligne de comptage. Nous considérons ci-après que la ligne de comptage est verticale (avec une orientation de $\pi/2$ ou $-\pi/2$). La direction du blob est obtenue en comparant son orientation avec les valeurs 0 (pour la droite) et $\pi$ (pour la gauche). Si son orientation est plus proche de 0 que de $\pi$, ça veut dire que le blob se dirige vers la droite et vice-versa. Nous calculons donc la valeur suivante :

$$\arg \max_{\theta \in \{0,\pi\}} (cos(B(i).O - \theta)) \tag{5.10}$$

où $B(i).O$ est l'orientation du blob $i$ et arg max est la fonction qui retourne l'ensemble des valeurs parmi l'ensemble $0, \pi$ qui maximise $cos(B(i).O - \theta)$. Si l'ensemble retourné est égal à $\{0\}$, alors le blob se dirige vers la droite. Si l'ensemble est égal à $\{\pi\}$, le blob se dirige vers la

gauche. Lorsque l'ensemble est égal à $\{0, \pi\}$, l'orientation du blob est égale à $\pi/2$ ou $-\pi/2$, ce qui correspond à un blob se déplaçant le long de la ligne de comptage. Dans ce dernier cas, le blob n'est pas pris en compte dans le comptage final puisqu'il se situe sur la ligne de comptage sans la franchir. Dans le cas d'une ligne de comptage horizontale, l'Équation 5.10 sera remplacée par :

$$\arg \max_{\theta \in \{\pi/2, -\pi/2\}} (cos(B(i).O - \theta)) \tag{5.11}$$

où $\pi/2$ représente la direction haut et $-\pi/2$ la direction bas. Cette méthode peut facilement être généralisée à n'importe quelle orientation de la ligne de comptage.

Une fois la direction du blob obtenue, le compteur final associé à cette direction est incrémenté par le nombre de personnes dans ce blob.

### 5.3.3 Expériences et résultats

Notre approche a été testée sur deux bases de vidéos correspondant à des configurations différentes. Nous avons attribué un identifiant à chaque vidéo de la façon suivante :

- Les vidéos 1, 2, 3, 4, 5, 6 proviennent de la première base correspondant à une configuration oblique. La durée totale des séquences est de 30 minutes à 25 images/seconde pour une résolution de $640 \times 240$ pixels.
- Les vidéos 7, 8, 9, 10, 11 proviennent de la deuxième base qui comprend des vidéos enregistrées par une caméra orientée verticalement. La durée totale des séquences est de 2 heures avec une cadence de 30 images/seconde pour une résolution de $240 \times 320$ pixels.

Le nombre de passants pour les deux ensembles dépasse les 1 000 personnes. Le Tableau 5.3 décrit brièvement chaque vidéo.

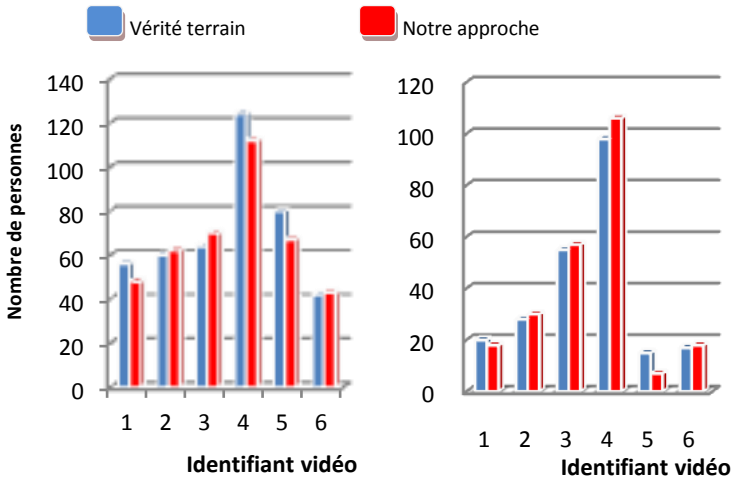
La première base de vidéos filme une voie piétonne traversée par des personnes et des voitures dans différentes conditions climatiques. La ligne virtuelle est placée verticalement au centre de la scène.

La seconde base de vidéos comprend des vidéos enregistrées par une caméra orientée verticalement. Pour les séquences 7 à 10, la caméra est placée à l'entrée d'un magasin dans le but d'estimer les flux des personnes entrantes et sortantes. La ligne virtuelle est placée horizontalement sur l'entrée. La séquence 11, quant à elle, filme l'entrée d'un laboratoire et la ligne virtuelle est placée verticalement sur l'entrée.

| Id | Configuration | Description |
|---|---|---|
| 1-2 | Oblique | Personnes passant individuellement |
| 3 | Oblique | Foule traversant la ruelle |
| 4-5 | Oblique | Présence de véhicules |
| 6 | Oblique | Temps pluvieux et personnes avec un parapluie |
| 7-10 | Verticale | Certaines personnes ont un caddie ou une poussette |
| 11 | Verticale | Cas de file indienne traversant dans les deux directions |

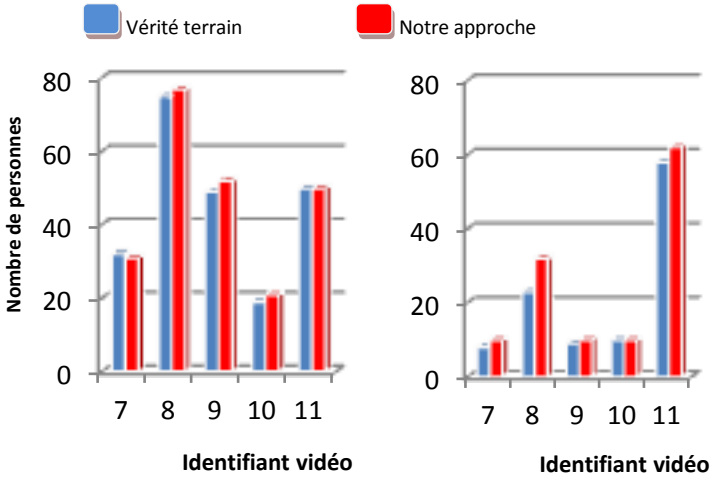
TABLE 5.3 – Description des ensembles de données

La Figure 5.20 présente les résultats de comptage lors de l'utilisation d'une caméra oblique, tandis que la Figure 5.21 présente ceux correspondants à une caméra verticale. Pour les deux figures, l'axe x correspond à l'identificateur de la séquence vidéo utilisé lors de l'évaluation, et l'axe y représente le nombre de personnes comptées. Les résultats relatifs à la vérité-terrain apparaissent en bleu, et ceux du comptage de notre système apparaissent en rouge.




(a) Gauche (b) Droite

FIGURE 5.20 – Résultats du comptage avec caméra oblique




(a) Entrée (b) Sortie

FIGURE 5.21 – Résultats du comptage avec caméra verticale

La Figure 5.22 représente les résultats obtenus pour l'intégralité d'une séquence vidéo issue de la base PETS2009. Cette figure représente une carte spatiotemporelle obtenue à partir de la ligne de comptage représentée en rouge dans la Figure 2.2. Chaque blob détecté est encadré par un rectangle contenant une flèche indiquant sa direction et un nombre représentant le comptage estimé. Le graphique montre la vérité-terrain et les résultats de notre approche pour chaque image.

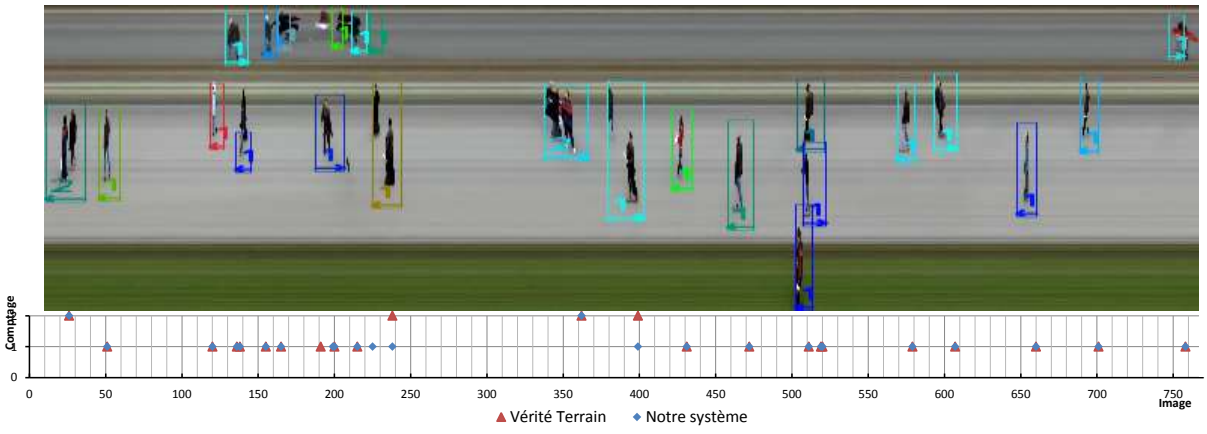



FIGURE 5.22 – Résultats du comptage dans une séquence vidéo issue de la base PETS2009

La précision globale du système pour un ensemble de séquences $Q$ et une certaine direction $d$ est définie par l'équation suivante :

$$A_{Q,d} = 1 - \frac{|\sum_{i \in Q} GT_{i,d} - \sum_{i \in Q} CS_{i,d}|}{\sum_{i \in Q} GT_{i,d}} \tag{5.12}$$

où

- $d$ est une valeur représentant la direction. $d = 0$ représente la direction "à gauche" ou les "entrées".
- $d = 1$ représente la direction "à droite" ou les "sorties".
- $d = 2$ représente les deux directions. $GT_{i,d}$ est la vérité terrain de la vidéo $i$ avec les directions $d$.
- $CS_{i,d}$ est le résultat du comptage effectué par notre système pour la vidéo $i$ avec la direction $d$.

Le Tableau 5.4 indique la précision de notre approche pour les séquences des deux bases de vidéos.

| Id | Direction | Vérité terrain | Notre système | Précision |
|---|---|---|---|---|
| 1-6 | (0) Gauche | 426 | 402 | 94.37% |
| | (1) Droite | 233 | 236 | 98.71% |
| | (2) Les deux | 659 | 638 | 96.81% |
| 7-11 | (0) Entrée | 225 | 231 | 97.33% |
| | (1) Sortie | 108 | 124 | 85.19% |
| | (2) Les deux | 333 | 335 | 93.39% |

TABLE 5.4 – Précision globale du système de comptage

Les résultats montrent que notre système s'avère d'une grande robustesse avec une précision globale de 96.81% pour le premier ensemble de vidéos, et de 93.39% pour le second. La grande différence entre $A_{7-11,0} = 97.33\%$ et $A_{7-11,1} = 85.19\%$ s'explique par le fait que les clients présents dans nos vidéos ont généralement rempli leurs sacs avant de quitter le magasin, ce qui génère des blobs plus grands. Nous remarquons également que le système a tendance à surcompter, mais la précision globale peut encore être améliorée par l'introduction d'un facteur

de correction. Cependant, les faux positifs sont davantage tolérés que les faux négatifs dans la plupart des applications. Dans le domaine sécuritaire, par exemple, il est préférable de rapporter des fausses détections plutôt que de manquer de réelles détections. Il en est de même dans le domaine du marketing où l'objectif est d'estimer le nombre de clients.

Les avantages de notre approche sont un temps de calcul faible et une précision élevée par rapport à des systèmes plus complexes. Notre algorithme a été implémenté sur une machine équipée d'un processeur Intel Celeron 1.86GHz CPU avec 1GB de mémoire vive sur laquelle nous pouvons traiter 45 images/seconde pour effectuer le comptage. La durée de traitement de notre approche est plus courte par rapport à d'autres autres approches qui fonctionnent à 12 images/seconde [ALCC09] (avec un processeur cadencé à 3.0GHz), à 25 images/seconde [CGZT09] (avec un processeur Pentium 4 2.8GHz CPU et 1GB de RAM) et à 33 images/seconde [VlTH06] (avec un processeur Pentium 2GHz).

## 5.4 Conclusion

Nous avons traité tout au long de ce chapitre trois problèmes de l'analyse du comportement humain dans une scène de foule, à savoir l'extraction des motifs de mouvement, la détection d'évènements et l'estimation des flux. Nous avons suivi une démarche à trois niveaux où le niveau bas correspond à l'estimation du flux optique. Le niveau intermédiaire, quant à lui, permet d'estimer le modèle directionnel pour les deux premiers problèmes et les blobs franchissant une ligne virtuelle pour le troisième problème. Le dernier niveau correspond à la mise en œuvre d'algorithmes d'apprentissage automatique et de classification afin d'atteindre les résultats escomptés.

Notre approche représente une contribution positive pour la détection des motifs de mouvement dans un environnement complexe. Ceci est possible grâce à la construction d'un modèle probabiliste associé aux orientations du mouvement dans une scène, et permet de distinguer des modalités de mouvements multiples à chaque localisation spatiale de la scène. Nos descripteurs et modèles contribuent également à la détection d'évènements se produisant au sein d'une

foule car ils permettent de suivre des groupes de personnes et d'extraire les caractéristiques des comportements de foule. De plus, notre méthode est capable de détecter un nombre d'évènements sémantiques important contrairement aux méthodes préalablement citées dans l'état de l'art (voir Section 2.6.2) qui se limitent aux situations anormales ou inhabituelles.

En ce qui concerne l'estimation des flux, notre approche s'inscrit dans la catégorie des approches spatiotemporelles. Elle a la particularité d'éviter la détection erronée de personnes s'arrêtant sur la ligne virtuelle. Pour ce faire, nous prenons en compte uniquement les pixels de la ligne virtuelle qui ont une magnitude de flux optique différente de zéro. Les vecteurs de flux optique indiquent également l'orientation des blobs de manière efficace. Par ailleurs, l'approche proposée a recours aux caractéristiques des blobs (vitesse de déplacement, orientation, position et dimensions) afin d'améliorer la précision en fonction du nombre de personnes se trouvant dans un blob. Notre approche est capable de détecter en temps réel les blobs franchissant la ligne virtuelle, contrairement aux approches précédentes proposées par [AMN01] et [BMB08], où il faut attendre que la carte spatiotemporelle soit totalement construite avant d'entamer la détection des blobs.

# Chapitre 6

# Conclusions et perspectives

## 6.1 Introduction

Ce mémoire de thèse a présenté différentes approches de vision par ordinateur pour analyser le comportement humain à partir de la vidéo en se basant sur l'orientation du mouvement. Les environnements ciblés peuvent être en intérieur ou en extérieur et la scène peut être individuelle (contient une seule personne) ou une scène de foule (contient un nombre important de personnes). Ce chapitre résume nos principales contributions ainsi que nos travaux futurs.

## 6.2 Résumé de nos contributions

Nous avons développé des méthodes suivant une démarche pyramidale en trois niveaux. Le premier niveau a pour but d'extraire des descripteurs de bas niveau depuis les images constituant une vidéo. Le niveau intermédiaire a pour but de calculer des descripteurs de niveau intermédiaire avec plus de sémantique. Le dernier niveau a pour but d'extraire des informations utiles aux utilisateurs. Nous avons traité quatre problèmes liés à l'analyse du comportement humain en suivant cette méthodologie. Nous résumons les principales contributions pour chaque problème.

***Analyse du comportement dans des scènes individuelles :*** Nous avons présenté un système de reconnaissance d'actions performant qui se base sur les modèles de direction et les modèles de magnitude du mouvement. Nous avons extrait les vecteurs de flux optique des séquences vidéo pour estimer les modèles cités plus haut. Le résultat est un modèle de séquence vidéo qui synthétise les principales orientations et magnitudes dans tous les blocs de la scène. Nous avons utilisé une mesure de distance pour détecter une action en comparant le modèle d'une séquence à des modèles de référence. En s'appuyant sur l'orientation et la magnitude du mouvement, notre approche aboutit à des résultats prometteurs comparés à d'autres approches de l'état de l'art, notamment sur des vidéos en haute définition.

***Analyse du comportement dans des scènes de foule :*** Nous avons analysé le comportement humain dans des scènes de foule à travers trois axes : l'extraction des motifs de mouvement, la détection des évènements et l'estimation des flux.

a) *Extraction des motifs de mouvement :* Nous avons présenté une nouvelle approche pour l'extraction de motifs de mouvements fondée sur l'analyse de l'orientation du mouvement dans une vidéo. L'approche extrait les vecteurs de flux optique et applique un algorithme adaptatif de regroupement sur les orientations du mouvement pour estimer des mélanges de lois de von Mises. Le modèle directionnel est ensuite estimé et contient les principales orientations à chaque région de la scène. Un algorithme de classification hiérarchique ascendante est appliqué sur le modèle de directionnel afin de détecter les motifs de mouvement. Les résultats expérimentaux sur plusieurs vidéos montrent l'efficacité de l'approche proposée dans les scènes complexes.

b) *Détection des évènements de foule :* Nous avons proposé une approche permettant de détecter des évènements dans des scènes de foule grâce à l'utilisation de modèles statistiques basés sur la magnitude et la vitesse du mouvement. Notre approche ne requiert pas la détection de chaque personne distinctement. En revanche, elle utilise des informations sur le mouvement global de façon à détecter des groupes de personnes ayant une même vitesse et une même orientation de mouvement.

   Deux catégories d'évènements sont ciblées : (i) les évènements liés à la vitesse qui incluent la marche et la course, et (ii) le rassemblement et la séparation qui incluent les évènements : rassemblement, séparation, dispersion locale et évacuation. Nous avons décrit chaque évènement et proposé un ensemble de caractéristiques relatives au comportement d'une foule par rapport à ces évènements. Nous avons ensuite mis en place deux classificateurs pour détecter les évènements dans chaque image. Les expériences réalisées à partir de la base PETS'2009, montrent que notre méthode est très prometteuse lorsqu'elle est appliquée aux scènes de foule contenant divers évènements.

c) *Estimation des flux :* Nous avons présenté une nouvelle approche permettant de compter le nombre d'individus franchissant une ligne de comptage. Les principales étapes de notre

méthodologie portent sur l'extraction des blobs franchissant cette ligne et l'estimation du nombre d'individus associés à ces blobs. L'approche proposée comporte trois originalités essentielles : (i) elle évite la détection des personnes statiques se trouvant sur la ligne de comptage, (ii) elle a recours à un algorithme de détection en ligne des blobs, et (iii) elle s'adapte facilement à différentes configurations de caméra grâce à l'utilisation d'un modèle de régression linéaire différent selon la configuration. Nous avons testé notre approche sur un vaste ensemble de vidéos contenant des vidéos réelles issues d'environnements extérieurs et intérieurs. Le système atteint une précision globale de 96.81% dans le cas de vidéos enregistrées par une caméra orientée de manière oblique, et de 93.39% dans le cas de vidéos enregistrées par une caméra orientée de manière verticale. Notre système est aussi performant pour gérer plusieurs lignes de comptage de manière robuste et efficace.

## 6.3 Travaux futurs

Bien que nous ayons effectué des recherches approfondies pour analyser le comportement humain depuis la vidéo, d'autres investigations peuvent être menées dans chacun des 4 problèmes que nous avons abordés. Nous proposons dans ce qui suit les orientations de recherche futures pour chaque problème.

Au niveau de l'analyse du comportement dans des scènes individuelles, nous avons traité le problème de reconnaissance d'actions. Nos travaux futurs s'orienteront vers l'amélioration de la flexibilité de notre approche par rapport à l'ajout ou la suppression de classes d'action et la reconnaissance d'actions dans des applications temps réel.

En ce qui concerne l'extraction des motifs de mouvement dans une scène de foule, une première investigation future consiste à appliquer l'approche proposée pour détecter des évènements anormaux. Ils seront considérés comme tout mouvement ne respectant pas le modèle directionnel. Une deuxième investigation future consiste à améliorer les performances des algorithmes de suivi en se servant du modèle directionnel comme une connaissance préalable du mouvement possible des personnes.

L'approche que nous avons proposée pour la détection des évènements de foule confond les ombres avec des personnes. Nous appliquerons une méthode permettant d'annihiler les effets d'ombre. Nous prendrons également en compte des informations de profondeur en 2,5D et en 3D pour mesurer les distances spatiales de manière plus précise.

Pour la partie estimation des flux, nous nous intéresserons à l'introduction d'un facteur de correction et à la création d'une base de vidéos utilisable par la communauté pour évaluer les systèmes de comptage pour des scénarios complexes.

L'ensemble de ces travaux sera publié sous forme d'une librairie open source pour l'analyse du comportement humain. Elle se présente sous forme d'un ensemble de fichiers source écrits en langage C/C++. Une application de démonstration des principales contributions de ce travail de thèse est aussi incluse. La Figure 6.1 illustre l'application de démonstration qui présente un ensemble de fenêtres qui affichent les résultats du calcul des descripteurs développés dans ce mémoire. La figure montre notamment un modèle directionnel en haut à droite et la détection des groupes de personnes en haut au milieu. Cette interface permet aussi l'affichage de divers résultats intermédiaires ainsi que la modification des paramètres de certains algorithmes.

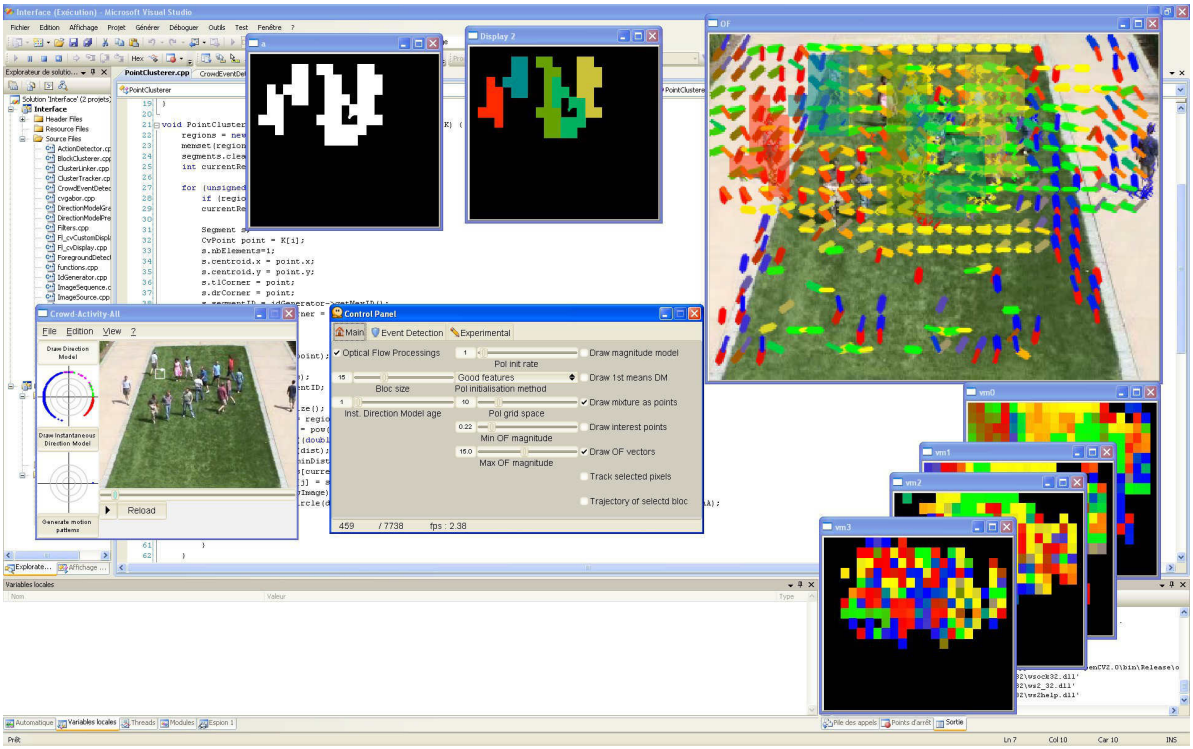



FIGURE 6.1 – Aperçu de notre application.

# Chapitre 7

# Publications

## 7.1 Livres et revues

[BID11] Yassine Benabbas, Nacim Ihaddadene et Chabane Djeraba. Motion pattern extraction and event detection for automatic visual surveillance. EURASIP Journal on Image and Video Processing, 2011 :15, 2011

[DLB10] Chaabane Djeraba, Adel Lablack et Yassine Benabbas. Multi-Modal User Interactions in Controlled Environments. Springer Publishing Company, 1st edition, 2010

## 7.2 Chapitres de livres

[BIY+10] Yassine Benabbas, Nacim Ihaddadene, Tarek Yahiaoui, Thierry Urruty et Chabane Dejraba. Event Detection in Crowd Scenes Using Statistical Models, chapter Advances in Knowledge Discovery and Management Vol. 2 (AKDM-2). Springer Publishing Company, 2010

## 7.3 Conférences Internationales

[BALD11] Yassine Benabbas, Samir Amir, Adel Lablack et Chabane Djeraba. Human action recognition using direction and magnitude models of motion. Internatioal Conference on Computer Vision and Applications (VISAPP), 2011

[BIYD10] Yassine Benabbas, Nacim Ihaddadene, Tarek Yahiaoui et Chabane Djeraba. Spatio-temporal optical flow analysis for people counting. International Conference on Advanced Video and Signal-Based Surveillance (AVSS), 2010

[BLID10] Yassine Benabbas, Adel Lablack, Nacim Ihaddadene et Chabane Djeraba. Action recognition using direction models of motion. International Conference on Pattern Recognition (ICPR), pages 4295-4298, 2010

[BID09] Yassine Benabbas, Nacim Ihaddadene et Chabane Djeraba. Global analysis of optical flow vectors for event detection in crowd scenes. International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), 2009

## 7.4 Conférences Nationales

[BYUD11] Yassine Benabbas, Tarek Yahiaoui, Thierry Urruty et Chabane Djeraba. Analyse spatiotemporelle des vecteurs de mouvement : application au comptage des personnes. 11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC), 2011

[BLUD11] Yassine Benabbas, Adel Lablack, Thierry Urruty et Chabane Djeraba. Reconnaissance d'actions par modélisation du mouvement. 11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC), 2011

[BIUD10] Yassine Benabbas, Nacim Ihaddadene, Thierry Urruty et Chabane Djeraba. Analyse globale du flux optique pour la détection d'évènements dans une scène de foule. Extraction et gestion des connaissances (EGC'2010), pages 339-350

[LUBD10] Adel Lablack, Thierry Urruty, Yassine Benabbas et Chabane Djeraba. Extraction de la région d'intérêt d'une personne sur un obstacle. Analyse globale du flux optique pour la détection d'évènements dans une scène de foule. Extraction et gestion des connaissances (EGC'2010), pages 683-684

# Bibliographie


[ABF06] Ernesto L. ANDRADE, Scott BLUNSDEN et Robert B. FISHER : Hidden markov models for optical flow analysis in crowds. *18th International Conference on Pattern Recognition. ICPR'06*, 1:460–463, 2006.

[AC97] J. K. AGGARWAL et Q. CAI : Human motion analysis : a review. *In IEEE Workshop on Nonrigid and Articulated Motion*, pages 90–102, 1997.

[ALCC09] B. ANTIC, D. LETIC, D. CULIBRK et V. CRNOJEVIC : K-means based segmentation for real-time zenithal people counting. *In International Conference on Image Processing (ICIP)*, 2009.

[AMN01] A. ALBIOL, I. MORA et V. NARANJO : Real-time high density people counter using morphological tools. *IEEE Conference on Intelligent Transportation Systems, Spain*, 2001.

[Ana89] P. ANANDAN : A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision (IJCV)*, 2:283–310, 1989. 10.1007/BF00158167.

[AQMM08] Catherine ACHARD, Xingtai QU, Arash MOKHBER et Maurice MILGRAM : A novel approach for recognition of human actions with semi-global features. *Machine Vision and Applications (MVA)*, 19:27–34, January 2008.

[ARSR08] A. ADAM, E. RIVLIN, I. SHIMSHONI et D. REINITZ : Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(3):555–560, March 2008.

[BALD11] Yassine BENABBAS, Samir AMIR, Adel LABLACK et Chabane DJERABA : Human action recognition using direction and magnitude models of motion. *In Internatioal Conference on Computer Vision and Applications (VISAPP)*, 2011.

[BB95] S. S. BEAUCHEMIN et J. L. BARRON : The computation of optical flow. *ACM Computer Survays*, 27:433–466, September 1995.

[BBE+08] Axel BAUMANN, Marco BOLTZ, Julia EBLING, Matthias KOENIG, Hartmut S. LOOS, Marcel MERKEL, Wolfgang NIEM, Jan Karl WARZELHAN, et Jie YU : A review and comparison of measures for automatic video surveillance systems. *EURASIP Journal on Image and Video Processing*, 2008.

[BCS07] Massimiliano BOZZOLI, Luigi CINQUE et Enver SANGINETO : A statistical method for people counting in crowded environments. *In 14th International Conference on Image Analysis and Processing (ICIAP)*, pages 506–511, 2007.

[BD01] A. F. BOBICK et J. W. DAVIS : The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(3):257–267, 2001.

[BEBV08] Thierry BOUWMANS, Fida El BAF et Bertrand VACHON : Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey. *Recent Patents on Computer Science*, 1(3):219–237, 11 2008.

[BETVG08] Herbert BAY, Andreas ESS, Tinne TUYTELAARS et Luc VAN GOOL : Speeded-up robust features (surf). *Computer Vision Image Understanding (CVIU)*, 110:346–359, June 2008.

[Bey00] D. BEYMER : Person counting using stereo. *In Workshop on Human Motion (HUMO)*, pages 127–133, 2000.

[BGS+05] Moshe BLANK, Lena GORELICK, Eli SHECHTMAN, Michal IRANI et Ronen BASRI : Actions as space-time shapes. *IEEE International Conference on Computer Vision (ICCV)*, 2:1395–1402, octobre 2005.

[BGS08] A. BASHARAT, A. GRITAI et M. SHAH : Learning object motion patterns for anomaly detection and improved object detection. *In International Conference on Computer Vision and pattern recognition (CVPR)*, pages 1–8, 2008.

[BGV92] Bernhard E. BOSER, Isabelle M. GUYON et Vladimir N. VAPNIK : A training algorithm for optimal margin classifiers. *In 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.

[BHQ+09] Henry BRAUN, Rafael HOCEVAR, Rossana B. QUEIROZ, Marcelo COHEN, Juliano Lucas MOREIRA, Julio C. JACQUES JÚNIOR, Adriana BRAUN, Soraia R. MUSSE et Ramini SAMADANI : Vhcve : A collaborative virtual environment including facial animation and computer vision. *In Proceedings of the 2009 VIII Brazilian Symposium on Games and Digital Entertainment*, SBGAMES '09, pages 207–213, Washington, DC, USA, 2009. IEEE Computer Society.

[BID09] Yassine BENABBAS, Nacim IHADDADENE et Chabane DJERABA : Global analysis of optical flow vectors for event detection in crowd scenes. *In International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2009.

[BID11] Yassine BENABBAS, Nacim IHADDADENE et Chabane DJERABA : Motion pattern extraction and event detection for automatic visual surveillance. *EURASIP Journal on Image and Video Processing*, 2011:15, 2011.

[Bir96] S. BIRCHFIELD : Klt : An implementation of the kanade-lucas-tomasi feature tracker, 1996.

[BIUD10] Yassine BENABBAS, Nacim IHADDADENE, Thierry URRUTY et Chabane DJERABA : Analyse globale du flux optique pour la détection d'évènements dans une scène de foule. *In Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC)*, pages 339–350, 2010.

[BIY+10] Yassine BENABBAS, Nacim IHADDADENE, Tarek YAHIAOUI, Thierry URRUTY et Chabane DEJRABA. : *Event Detection in Crowd Scenes Using Statistical Models*, chapitre Advances in Knowledge Discovery and Management Vol. 2 (AKDM-2). Springer Publishing Company, 2010.

[BIYD10] Yassine BENABBAS, Nacim IHADDADENE, Tarek YAHIAOUI et Chabane DJERABA : Spatio-temporal optical flow analysis for people counting. *In International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2010.

[BLID10] Yassine BENABBAS, Adel LABLACK, Nacim IHADDADENE et Chabane DJERABA : Action recognition using direction models of motion. *In International Conference on Pattern Recognition (ICPR)*, pages 4295–4298, 2010.

[BLUD11] Yassine BENABBAS, Adel LABLACK, Thierry URRUTY et Chabane DJERABA : Reconnaissance d'actions par mod'elisation du mouvement. *In 11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC)*, 2011.

[BM04] S. BAKER et I. MATTHEWS : Lucas-kanade 20 years on : A unifying framework. *International Journal of Computer Vision (IJCV)*, 56(3):221–225, 2004.

[BMB08] J. BARANDIARAN, B. MURGUIA et Fe. BOTO : Real-time people counting using multiple lines. *IEEE Ninth International Workshop on Image Analysis for Multimedia Interactive Services, Klagenfurt, Austria*, 2008.

[BMV09] Antoni B.CHAN, Mulloy MORROW et Nuno VASCONCELOS : Analysis of crowded scenes using holistic properties. *In 11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2009.

[Bob97] Aaron F. BOBICK : Movement, activity and action : The role of knowledge in the perception of motion. *Philosophical Transactions : Biological Sciences*, 352(1358):1257–1265, 1997.

[BV99] B.A. BOGHOSSIAN et S.A. VELASTIN : Motion-based machine vision techniques for the management of large crowds. *In Proceedings of ICECS '99. The 6th IEEE International Conference on Electronics, Circuits and Systems*, volume 2, pages 961–964, 1999.

[BYUD11] Yassine BENABBAS, Tarek YAHIAOUI, Thierry URRUTY et Chabane DJERABA : Analyse spatiotemporelle des vecteurs de mouvement : application au comptage

des personnes. *In Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC)*, pages 173–178, 2011.

[CABT04] F. CUPILLARD, A. AVANZI, F. BREMOND et M. THONNAT : Video understanding for metro surveillance. *2004 IEEE International Conference on Networking, Sensing and Control*, 1:186–191, 2004.

[CGZT09] Yang CONG, Haifeng GONG, Song-Chun ZHU et Yandong TANG : Flow mosaicking : Real-time pedestrian counting without scene-specific learning. *In International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1093–1100, 2009.

[CK05] S. CHEUNG et C. KAMATH : Robust background subtraction with foreground validation for urban traffic video. *EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Intelligent Vision Systems : Methods and Applications*, 14:2330–2340, 2005.

[CLB04] Chao CHEN, Andy LIAW et Leo BREIMAN : Using Random Forest to Learn Imbalanced Data. Rapport technique, Department of Statistics, University of Berkeley, 2004.

[CLK00] Robert T. COLLINS, Alan J. LIPTON et Takeo KANADE : Introduction to the special section on video surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(8):745–746, 2000.

[DLB10] Chaabane DJERABA, Adel LABLACK et Yassine BENABBAS : *Multi-Modal User Interactions in Controlled Environments*. Springer Publishing Company, 1st édition, 2010.

[DLR77] A.P. DEMPSTER, N.M. LAIRD et D.B. RUBIN : Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[DRC+05] P. DOLLAR, V. RABAUD, G. COTTRELL, et S. BELONGIE : Behavior recognition via sparse spatio-temporal features. *In International Workshop on Visual Sur-*

*veillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, 2005.

[DRCB05] P. DOLLAR, V. RABAUD, G. COTTRELL et S. BELONGIE : Behavior recognition via sparse spatio-temporal features. *In 2nd International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (PETS)*, pages 65–72, 2005.

[DYV95] A.C. DAVIES, Jia Hong YIN et S.A. VELASTIN : Crowd monitoring using image processing. *Electronics & Communication Engineering Journal*, 7(1):37–47, 1995.

[EBBV07] Fida El BAF, Thierry BOUWMANS et Bertrand VACHON : Comparison of background subtraction methods for a multimedia learning space. *In International Conference on Signal Processing and Multimedia (SIGMAP)*, Barcelona - Spain, July 2007.

[EESA08] S. ELHABIAN, K. EL-SAYED et S. AHMED : Moving object detection in spatial domain using background removal techniques - state-of-art. *Recent Patents on Computer Science*, 1(1):32–54, 2008.

[EG09] Markus ENZWEILER et Dariu M. GAVRILA : Monocular pedestrian detection : Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(12):2179–2195, 2009.

[FAI+05] David A. FORSYTH, Okan ARIKAN, Leslie IKEMOTO, James O'BRIEN et Deva RAMANAN : Computational studies of human motion : part 1, tracking and motion synthesis. *Found. Trends. Comput. Graph. Vis.*, 1(2-3):77–254, 2005.

[FR97] N. FRIEDMAN et S. RUSSELL : Image segmentation in video sequences : A probabilistic approach. *In 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 175–181, 1997.

[FS95] Yoav FREUND et Robert E. SCHAPIRE : A decision-theoretic generalization of on-line learning and an application to boosting. *In European Conference on Computational Learning Theory*, pages 23–37, 1995.

[Gav99] Dariu M. GAVRILA : The visual analysis of human movement : a survey. *Computer Vision and Image Understanding (CVIU)*, 1(73):82–92, 1999.

[GB80] Gary L. GAILE et James E. BURT : *Directional Statistics*. Concepts and techniques in modern geography. Norwich, England : Geo Abstracts, 1980.

[GBJ+07] A. GARDEL, I. BRAVO, P. JIMENEZ, J.L. LAZARO et A. TORQUEMADA : Real time head detection for embedded vision modules. *In IEEE International Symposium on Intelligent Signal Processing (WISP)*, pages 1–6, 2007.

[GBS+07] Lena GORELICK, Moshe BLANK, Eli SHECHTMAN, Michal IRANI et Ronen BASRI : Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(12):2247–2253, 2007.

[GBTT08] G.GARCIA-BUNSTER et M. TORRES-TORRITI : Effective pedestrian detection and counting at bus stops. *In IEEE Latin American Robotic Symposium*, pages 158–163, 2008.

[GK95] Lynne GREWE et Avinash C. KAK : Interactive learning of a multiple-attribute hash table classifier for fast object recognition. *Computer Vision and Image Understanding (CVIU)*, 61(3):387–416, 1995.

[GLSG10] David GERONIMO, Antonio M. LOPEZ, Angel D. SAPPA et Thorsten GRAF : Survey of pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(7):1239–1258, juillet 2010.

[GT07] T. GANDHI et M. M. TRIVEDI : Pedestrian protection systems : Issues, survey, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 8(3): 413–430, septembre 2007.

[GWT09] Jacob M. GRYN, Richard P. WILDES et John K. TSOTSOS : Detecting motion patterns via direction maps with application to surveillance. *Computer Vision and Image Understing*, 113(2):291–307, 2009.

[HAS08a] Min HU, Saad ALI et Mubarak SHAH : Detecting global motion patterns in complex videos. *In ICPR'08 : International Conference on Pattern Recognition*, 2008.

[HAS08b] Min HU, Saad ALI et Mubarak SHAH : Learning motion patterns in crowded scenes using motion flow field. *In International Conference on Pattern Recognition (ICPR)*, 2008.

[HS81a] Berthold K. P. HORN et Brian G. SCHUNCK : Determining optical flow. *ARTIFICAL INTELLIGENCE*, 17:185–203, 1981.

[HS81b] B.K.P. HORN et B.G. SCHUNK : Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[HS88] C. HARRIS et M.J. STEPHENS : A combined corner and edge detector. *In Alvey Vision Conference*, pages 147–152, 1988.

[HTWM04] Weiming HU, Tieniu TAN, Liang WANG et S. MAYBANK : A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C : Applications and Reviews*, 34:334–352, 2004.

[HXF+06] Weiming HU, Xuejuan XIAO, Zhouyu FU, Dan XIE, Tieniu TAN et Steve MAYBANK : A system for learning statistical motion patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(9):1450–1464, 2006.

[ID08] Nacim IHADDADENE et Chabane DJERABA : Real-time crowd motion analysis. *In International Conference on Pattern Recognition (ICPR)*, 2008.

[JBEJ94] Gunnar JOHANSSON, Sten Sture BERGSTROM, William EPSTEIN et Gunnar JANSSON : *Perceiving Events and Objects*. Lawrence Erlbaum Associates, 1994.

[KB01] P. KAEWTRAKULPONG et R. BOWDEN : An improved adaptive background mixture model for realtime tracking with shadow detection. *In 2nd European Workshop on Advanced Video Based Surveillance Systems*, 2001.

[KKUG07] V. KRÜGER, D. KRAGIC, A. UDE et C. GEIB : The meaning of action : A review on action recognition and mapping. *Advanced Robotics*, 21(13):1473–1501, 2007.

[KLS03] Michael KOCKELKORN, Andreas LÜNEBURG et Tobias SCHEFFER : Using transduction and multi-view learning to answer emails. *In 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 266–277, Cavtat-Dubrovnik, Croatia, September 2003.

[KMS08] Alexander KLÄSER, Marcin MARSZAŁEK et Cordelia SCHMID : A spatio-temporal descriptor based on 3d-gradients. *In British Machine Vision Conference (BMVC)*, pages 995–1004, sep 2008.

[KN09] L. KRATZ et K. NISHINO : Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. *In International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1453, 2009.

[LCC01] Sheng-Fuu LIN, Jaw-Yeh CHEN et Hung-Xin CHAO : Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 31(6):645–654, 2001.

[LCSL07] Ivan LAPTEV, Barbara CAPUTO, Christian SCHÜLDT et Tony LINDEBERG : Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding (CVIU)*, 108:207–229, December 2007.

[LGF09] D. LIN, W.E.L. GRIMSON et J. FISHER : Learning visual flows : A lie algebraic approach. *In International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 747–754, 2009.

[LH02] B. LEE et M. HEDLEY : Background estimation for video surveillance. *In Image and Vision Computing New Zealand (IVCNZ)*, pages 315–320, 2002.

[LK81] B. D. LUCAS et T. KANADE : An iterative image registration technique with an application to stereo vision. *In International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679, 1981.

[LL03] Ivan LAPTEV et Tony LINDEBERG : Space-time interest points. *In International Conference on Computer Vision (ICCV)*, 2003.

[LL04] I. LAPTEV et T. LINDEBERG : Velocity adaptation of space-time interest points. *In International Conference on Pattern Recognition (ICPR)*, pages 52–56, 2004.

[LL06] Wei-Lwun LU et J. J. LITTLE : Simultaneous tracking and action recognition using the pca-hog descriptor. *In The 3rd Canadian Conference on Computer and Robot Vision*, page 6, 2006.

[LMSR08] Ivan LAPTEV, Marcin MARSZAŁEK, Cordelia SCHMID et Benjamin ROZENFELD : Learning realistic human actions from movies. *In International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[Low04] David G. LOWE : Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60:91–110, November 2004.

[LTJS12] Xiaowei LIU, Shasha TIAN, Jiafu JIANG et Jing SHEN : Moving human head detection for automatic passenger counting system. *In* Zhihong QIAN, Lei CAO, Weilian SU, Tingkai WANG et Huamin YANG, éditeurs : *Recent Advances in Computer Science and Information Engineering*, volume 125 de *Lecture Notes in Electrical Engineering*, pages 147–152. Springer Berlin Heidelberg, 2012.

[LTR+05] X. LIU, P. H. TU, J. RITTSCHER, A. PERERA et N. KRAHNSTOEVER : Detecting and counting people in surveillance applications. *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, NY, USA*, pages 306–311, 2005.

[LUBD10] Adel LABLACK, Thierry URRUTY, Yassine BENABBAS et Chabane DJERABA : Extraction de la région d'intérêt d'une personne sur un obstacle. *In Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC)*, pages 683–684, 2010.

[LVF03] A. LEVIN, P. VIOLA et Y. FREUND : Unsupervised improvement of visual detectors using co-training. *In International Conference on Computer Vision (ICCV)*, volume I, pages 626–633, 2003.

[LYSS12] Jingen LIU, Yang YANG, Imran SALEEMI et Mubarak SHAH : Learning semantic features for action recognition via diffusion maps. *Computer Vision and Image Understanding*, 116(3):361–377, 2012.

[MHK06] Thomas B. MOESLUND, Adrian HILTON et Volker KRUGER : A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, novembre 2006.

[MJD+00] Stephen J. MCKENNA, Sumer JABRI, Zoran DURIC, Azriel ROSENFELD et Harry WECHSLER : Tracking groups of people. *Computer Vision and Image Understanding*, 80(1):42–56, 2000.

[MLHT04] R. MA, L. LI, W. HUANG et Q. TIAN : On pixel count based crowd density estimation for visual surveillance. *IEEE Conference Cybernetics and Intelligent Systems*, 1:170–173, 2004.

[MMSZ05] S. MESSELODI, C. MODENA, N. SEGATA et M. ZANIN : A kalman filter based background updating algorithm robust to sharp illumination changes. *In 13th International Conference on Image Analysis and Processing (ICIAP)*, pages 163–170, Cagliari - Italy, 2005.

[MOS09] Ramin MEHRAN, Alexis OYAMA et Mubarak SHAH : Abnormal crowd behavior detection using social force model. *International Conference on Computer Vision and Pattern Recognition (CVRP)*, 2009.

[MPK09] Ross MESSING, Chris PAL et Henry KAUTZ : Activity recognition using the velocity histories of tracked keypoints. *In International Conference on Computer Vision (ICCV)*, 2009.

[MT08] B.T. MORRIS et M.M. TRIVEDI : A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8):1114 –1127, aug. 2008.

[Pet00] J. K. PETERSEN : *Understanding Surveillance Technologies*. CRC Press ; 1 edition, 2000.

[Pic04] M. PICCARDI : Background subtraction techniques : A review. *In International Conference on Systems, Man and Cybernetics (SMC)*, The Hague, The Netherlands, 2004.

[POP98] C.P. PAPAGEORGIOU, M. OREN et T. POGGIO : A general framework for object detection. *In International Conference on Computer Vision (ICCV)*, pages 555–562, 1998.

[Pop10] Ronald POPPE : A survey on vision-based human action recognition. *Image and Vision Computing (IVC)*, 28(6):976 – 990, 2010.

[Por03] F. PORIKLI : Human body tracking by adaptive background models and mean-shiftanalysis. *In IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2003)*, 2003.

[PT05] F. PORIKLI et O. TUZEL : Bayesian background modeling for foreground detection. *In ACM Int Workshop on Video Surveillance and Sensor Networks (VSSN)*, pages 55–58, November 2005.

[PVM07] A. PANDE, A. VERMA et A. MITTAL : Network aware optimal resource allocation for e-learning videos. *In 6th International Conference on mobile Learning*, Melbourne - Australia, 2007.

[RAK09] Mikel RODRIGUEZ, Saad ALI et Takeo KANADE : Tracking in unstructured crowded scenes. *In International Conference on Computer Vision (ICCV)*, Kyoto, 2009.

[RBK96] Henry A. ROWLEY, Shumeet BALUJA et Takeo KANADE : Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20:23–38, 1996.

[Ric11] Szeliski RICHARD : *Computer Vision Algorithms and Applications*. Springer, 2011.

[SaMCC08] Mei-Ling SHYU, Zongxing Xie abd MIN CHEN et Shu-Ching CHEN : Video semantic event/concept detection using a subspace-based multimedia datamining framework. *IEEE transactions on multimedia ISSN 1520-9210*, 10:252–259, 2008.

[SAS07] Paul SCOVANNER, Saad ALI et Mubarak SHAH : A 3-dimensional sift descriptor and its application to action recognition. *In Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 357–360, New York, NY, USA, 2007. ACM.

[SG99] C. STAUFFER et W.E.L. GRIMSON : Adaptative background mixture models for real-time tracking. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 246–252, 1999.

[SG00] Chris STAUFFER et W. Eric L. GRIMSON : Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(8):747–757, 2000.

[SLBS06] O. SIDLA, Y. LYPETSKYY, N. BRANDLE et S. SEER : Pedestrian detection and tracking for counting applications in crowded situations. *EEE International Conference on Video and Signal Based Surveillance (AVSS 2006), IEEE Computer Society , Washington, DC, USA*, pages 70–75, 2006.

[SLC04] Christian SCHULDT, Ivan LAPTEV et Barbara CAPUTO : Recognizing human actions : A local svm approach. *In International Conference on Pattern Recognition (ICPR)*, 2004.

[SMB00] Cordelia SCHMID, Roger MOHR et Christian BAUCKHAGE : Evaluation of interest point detectors. *International Journal of Computer Vision (IJCV)*, 37(2):151–172, 2000.

[SMP08] M. SIGARI, N. MOZAYANI et H. POURREZA : Fuzzy running average and fuzzy background subtraction : concepts and application. *Int J Comput Sci Network Security*, 8(2):138–143, 2008.

[SSJNFF08] Andrey DelPozo SILVIO SAVARESE, JUAN, Carlos NIEBLES et Li FEI-FEI : Spatial-temporal correlations for unsupervised action classification. *In Proceedings of the Workshop on Applications of Computer Vision (WACV)*, 2008.

[ST94] Jianbo SHI et Carlo TOMASI : Good features to track. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.

[Ste03] Robert STERNBERG : *Cognitive Psychology, Third Edition*. Thomson Wadsworth, 2003.

[TCSU08] P. TURAGA, R. CHELLAPPA, V. S. SUBRAHMANIAN et O. UDREA : Machine recognition of human activities : A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.

[TH08] Christian THURAU et Vaclav HLAVAC : Pose primitive based human action recognition in videos or still images. *In International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, juin 2008.

[TKBM99] K. TOYAMA, J. KRUMM, B. BRUMITT et B. MEYERS : Wallflower : Principles and practice of background maintenance. *In International Conference on Computer Vision (ICCV)*, pages 255–261, Corfu - Greece, 1999.

[Tri04] Bill TRIGGS : Detecting keypoints with stable position, orientation and scale under illumination changes. *In* Tomáš PAJDLA et Jiří MATAS, éditeurs : *8th European Conference on Computer Vision (ECCV)*, volume 3024 de *Lecture Notes in Computer Science*, pages 100–113, Prague, Tchéquie, mai 2004. Springer.

[TYOY99] K. TERADA, D. YOSHIDA, S. OE et J. YAMAGUSHI : A counting method of the number of passing people using a stereo camera. *IEEE 25th Annual Conference of Industrial Electronics Society, San Jose, California, USA*, pages 338–342, 1999.

[UKS09] A. UTASI, A. KISS et T. SZIRÁNYI : Statistical filters for crowd image analysis. *In Performance Evaluation of Tracking and Surveillance workshop at CVPR 2009*, pages 95–100, Miami, Florida, 2009.

[VJS03] Paul VIOLA, Michael JONES et Daniel SNOW : Detecting pedestrians using patterns of motion and appearance. *In International Conference on Computer Vision (ICCV)*, pages 734–741, 2003.

[VlTH06] Senem VELIPASALAR, Ying li TIAN et Arun HAMPAPUR : Automatic counting of interacting people by using a single uncalibrated camera. *In International Conference on Multimedia and Expo (ICME)*, pages 1265–1268, 2006.

[vOBK11] Tim van OOSTERHOUT, Sander BAKKES et Ben J. A. KRÖSE : Head detection in stereo data for people counting and segmentation. *In Sixth International Confe-*

*rence on Computer Vision Theory and Applications (VISAPP)*, pages 620–625, 2011.

[WADP97a] C. R. WREN, A. AZARBAYEJANI, T. DARRELL et A. P. PENTLAND : Pfinder : real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[WADP97b] Christopher Richard WREN, Ali AZARBAYEJANI, Trevor DARRELL et Alex PENTLAND : Pfinder : Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7):780–785, 1997.

[Wan00] Y. WANG : A new approach to fitting linear models in high dimensional spaces. *PhD thesis, Department of Computer Science, University of Waikato, New Zealand*, 2000.

[WHT03] Liang WANG, Weiming HU et Tieniu TAN : Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, mars 2003.

[WMS10] Shandong WU, Brian E. MOORE et Mubarak SHAH : Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. *In International Conference on Computer Vision and Pattern Recognition (CVPR)*, San Fransisco, California, june 2010.

[WS06] H. WANG et D. SUTER : A novel robust statistical method for background initialization and visual surveillance. *In Asian Conference on Computer Vision (ACCV)*, pages 328–337, Hyderabad - India, January 2006.

[WTG06] Xiaogang WANG, Kinh TIEU, et Eric GRIMSON : Learning semantic scene models by trajectory analysis. *In European Conference on Computer Vision (ECCV)*, 2006.

[WTG08] Geert WILLEMS, Tinne TUYTELAARS et Luc GOOL : An efficient dense and scale-invariant spatio-temporal interest point detector. *In Proceedings of the 10th European Conference on Computer Vision : Part II*, pages 650–663, Berlin, Heidelberg, 2008. Springer-Verlag.

[XWLZ07] XiaoWei. XU, ZhiYan WANG, YingHong LIANG et YanQing ZHANG : A rapid method for passing people counting in monocular video sequences. *The Sixth International Conference on Machine Learning and Cybernetics, Hong Kong*, pages 1657–1662, 2007.

[YCSX08] Shengsheng YU, Xiaoping CHEN, Weiping SUN et Deping XIE : A robust method for detecting and counting people. *International Conference on Audio, Language and Image Processing (ICALIP 2008), Iceland*, pages 1545–1549, 2008.

[YGBG03] Danny B. YANG, Hector H. GONZALEZ-BANOS et Leonidas J. GUIBAS : Counting people in crowds with a real-time network of simple image sensors. *In International Conference on Computer Vision (ICCV)*, pages 122–129, 2003.

[YK08] T. YAHIAOUI et L. KHOUDOUR : A people counting system based on dense and close stereovision. *IEEE International Conference on Image and Signal processing (ICISP 2008), Springer,Cherbourg France*, pages 59–66, 2008.

[YM09] Q. YU et G. MEDIONI : Motion pattern interpretation and detection for tracking moving vehicles in airborne video. *In International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2671–2678, 2009.

[YR06] Y.JEON et P. RYBSKI : Analysis of a spatio-temporal clustering algorithm for counting people in a meeting. *tech. report CMU-RI-TR-06-04, Robotics Institute, Carnegie Mellon University*, 2006.

[ZC07] E. ZHANG et F. CHEN : A fast and robust people counting method in video surveillance. *Int. Conf. on Computational Intelligence and Security, China*, pages 339–343, 2007.

[ZDC09] Xi ZHAO, Emmanuel DELLEANDREA et Liming CHEN : A people counting system based on face detection and tracking in a video. *In International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 67–72, 2009.

[ZLL09] T.Z. ZHANG, H.Q. LU et S.Z. LI : Learning semantic scene models by object classification and trajectory clustering. *In International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1940–1947, 2009.

[ZMR+08] Beibei ZHAN, Dorothy MONEKOSSO, Paolo REMAGNINO, Sergio VELASTIN et Li-Qun XU : Crowd analysis : a survey. *Machine Vision and Applications*, 19: 345–357, 2008.

[ZN03] Tao ZHAO et Ram NEVATIA : Bayesian human segmentation in crowded situations. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2:459, 2003.

[ZN04] Tao ZHAO et Ram NEVATIA : Tracking multiple humans in complex situations. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 26(9):1208–1221, 2004.

[ZWW05] Yanqing ZHANG, Zhiyan WANG et Bin WANG : A camera calibration method based on nonlinear model and improved planar pattern. *JCIS/CVPRIP*, 3:707–7012, 2005.

# Annexe A

# Définitions

Cette partie décrit brièvement les termes fréquemment utilisés dans ce document :

**Action** C'est une succession de mouvements simples effectués par une seule et même personne durant un laps de temps court (ex : sauter, tirer au but, effacer le tableau...). La figure A.1 montre des exemples d'actions de la vie quotidienne.

(a) (b)

FIGURE A.1 – Actions de la vie quotidienne, (a) écrire sur un tableau, (b) boire de l'eau

**Évènement** Un évènement peut avoir plusieurs significations. Dans le contexte de la vidéo-surveillance c'est toute action ou succession d'actions effectuées par une ou plusieurs

personnes comme illustré dans la figure A.2. On parle aussi d'évènement anormal ou suspect pour tout évènement qui peut nuire à la sécurité et à l'ordre général.

FIGURE A.2 – Évènement de foule correspondant à une évacuation d'urgence (mise en scène fictive)

**Caractéristique** ou descripteur. Ce sont des informations que l'on calcule à partir des pixels d'une image. Quelques exemples : bords (edges), histogramme (histogram), ...etc

**Flux ou flot** Mouvement cohérent de points (pixels) ou caractéristiques entre des images successives.

**Flux optique** C'est le mouvement apparent des objets d'une scène causé par le mouvement relatif entre un observateur (l'œil humain ou une camera) et la scène. Ce mouvement est extrait sous forme de vecteur comme illustré dans la figure A.3.

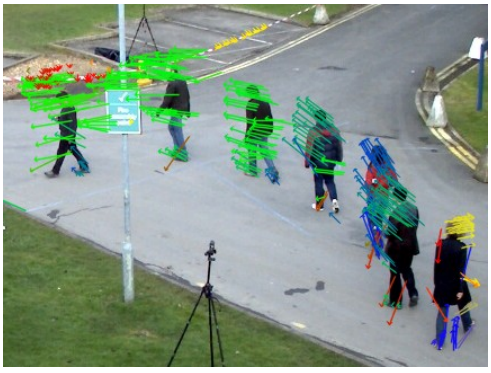


FIGURE A.3 – Les vecteurs de flux optique

**Histogramme** Permet d'avoir un résumé ou condensé de certaines informations d'une image. Par exemple un histogramme de couleurs donne pour chaque couleur le nombre de pixels ayant cette couleur, ainsi, on pourra connaître la couleur dominante d'une image.*

**Image binaire** C'est une image dans laquelle un pixel n'a que deux valeurs possibles qui sont 1 allumé ou éteint 0. D'où le nom binaire. Ce type d'images peut être traité très rapidement.

**Soustraction de l'arrière plan** ou soustraction du fond. Consiste à détecter les zones inutiles des images d'une vidéo qui sont appelées fond ou arrière plan. Les zones utiles sont appelées avant plan ou premier plan. Ces zones dépendent de l'application, par exemple dans le cas de la détection des personnes, l'arrière plan regroupe les voitures, le ciel, la route, etc.

**Statistique directionnelle** C'est une discipline des statistiques qui fournit des outils mathématiques pour traiter les observations angulaires ou celles qui sont définies à deux-pi près ; $\alpha = \alpha + 2k\pi, k \in \mathbb{Z}$. L'angle 0 correspond à l'axe x et les angles sont mesurés au sens inverse des aiguilles d'une montre par convention (voir $\theta_p$ sur la Figure A.4). Gaile et Burt [GB80] ont posé les premières bases et outils de cette discipline en 1980.

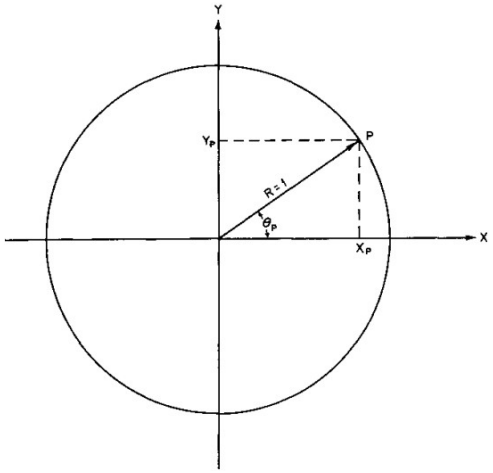



FIGURE A.4 – Illustration d'un cercle unitaire où on mesure l'angle $\theta_p$

# Annexe B

# Description de MIAUCE

Cette thèse a été réalisée dans le cadre du projet européen MIAUCE (Multi-modal Interactions Analysis and exploration of Users within a Controlled Environment). Nous avons contribué à la partie estimation des flux et à la participation à la compétition PETS 2009. Ci-dessous, quelques mots sur le projet MIAUCE.

Le projet européen MIAUCE vise à étudier et développer des techniques pour analyser le comportement multimodale des utilisateurs dans un contexte applicatif. Le comportement multimodale prend la forme de regard fixe, de clignotement de l'œil ou de mouvements de corps.

Le projet MIAUCE est un projet européen qui regroupe 8 partenaires : CNRS (Lille, France), Université d'Amsterdam (Pays-Bas), Université de Glasgow (Ecosse), Université de Trento (Italie), Université de Namur (Belgique), Sylis (Nantes, France), Visual Tools (Madrid, Espagne) et Tilde (Riga, Lettonie). Les différents partenaires du projet MIAUCE B.1 étudient et développent des techniques qui capturent et analysent le comportement multimodale dans les environnements contrôlés. En raison d'une telle analyse, l'information est adaptée aux besoins d'utilisateur et à la situation donnée. Ils étudient l'utilisation et l'efficacité de leur technique dans trois applications différentes telles que la sécurité, la vente adaptée aux besoins du client, et la TV interactive sur le Web. L'objectif est donc de développer des techniques d'analyse

d'interactions de l'homme dans un environnement contrôlé (supermarché, etc..), plutôt que des interactions de l'homme avec un ordinateur.

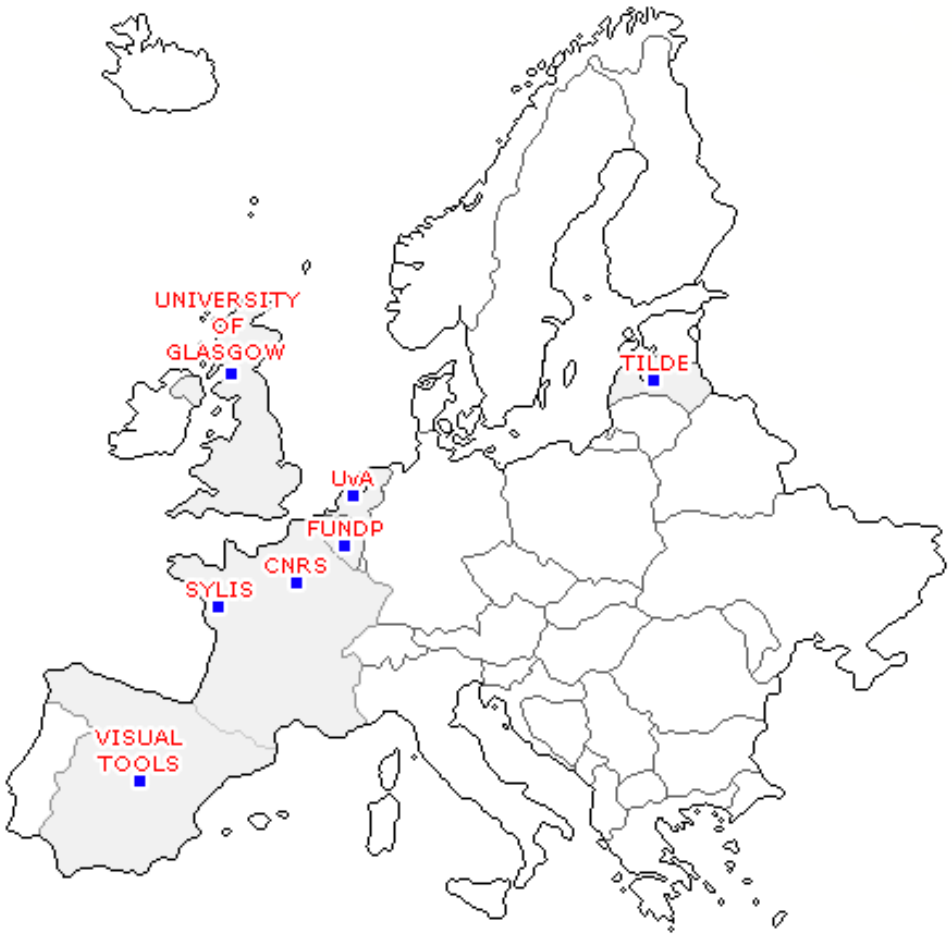



FIGURE B.1 – Les différents partenaires du projet MIAUCE

Les techniques sont alors développées et validées dans trois différents domaines d'applications. Ceci permettra de développer des techniques généralisables et d'ouvrir des voies pour l'exploitation industrielle. Les applications développées seront faites par le biais d'une aide industrielle étant donné que ce projet est réalisé pour des raisons markéting, mais aussi de sécurité. Dans ce contexte, le CRID se concentrera sur les problèmes légaux que de telles technologies impliquent, avec pour objectif d'adresser des recommandations sur le design du système d'information et sa gestion. L'aspect éthique et sociologique sera étudié au sein de la CITA. En conclusion, le projet étudie les conditions légales et morales nécessaires pour concevoir ces

nouvelles applications. En outre, l'acceptabilité éthique de ces nouvelles technologies d'intelligences que sont la capture multimodale de comportement, sera étudiée afin de garantir le succés de leur exécution.

L'objectif Principale des 3 partenaires industriels (Sylis, Visual Tools, Tilde) est de valider et expérimenter les technologies développées par les partenaires universitaires dans des applications déployées dans un environnement grandeur nature. Trois tâches qui correspondent à 3 différentes applications sont supportés par des industriels afin d'étudier la généralité de la technologie développée. Les applications ciblées par ce projet sont :

- Sécurité : La détection de personnes qui tombent à la sortie d'un escalier mécanique dans un aéroport.
- Marketing : L'estimation des produits regardé par les personnes qui passent devant une vitrine.
- Interactive WebTV : Proposer du contenu vidéo à une personne qui regarde une WebTV en se basant sur l'émotion.

# Annexe C

# Description du projet CAnADA

Dans le cadre d'une absence d'offre technologique en matière de détection en temps réel à partir de la vidéo des comportements anormaux de personnes dans un lieu public, tel un lieu de vente, des industriels comme YOUG'S et Thales expriment ce besoin.

Le but du projet CAnADA 'Comportements Anormaux : Analyse, Détection, Alerte' est de proposer une approche pour la détection en temps réel de comportements inhabituels pouvant mettre en péril la sécurité des personnes et des biens dans des lieux publics, comme les centres commerciaux, les magasins, les métros. Les informations détectées seront transmises à une application capable de rendre en temps réel une alarme et de ramener la situation à un niveau normal via un affichage par exemple. Dans ce cadre, un réseau de caméras est mis en place, comportant à certains endroits de la scène des zones aveugles dont il faudra tenir compte (un individu peut se cacher dans une telle zone afin de définir une stratégie de vol, hors des caméras). Les traitements mis en place consistent à extraire les trajectoires des personnes, ainsi que leurs activités, en tenant compte du contexte de la scène (disposition des caméras et des objets de la scène), et en traitant les cas d'occultations (une personne cachant une autre personne tout ou partie, ou bien un objet cachant un membre d'une personne). Les zones du visage des personnes suivies doivent être masquées, car il ne faut pas avoir accès à l'identité des personnes, le partenariat CNIL nous guidant dans cet aspect.

Plusieurs partenaires scientifiques, juridiques et industriels sont regroupés dans ce consortium, couvrant ainsi des compétences complémentaires :

- Le LIRIS (Laboratoire d'InfoRmatique en Images et Systèmes d'information), INSA de Lyon, et la société FOXTREAM, tous deux spécialiste dans l'analyse des objets en mouvement, la gestion des occultations, la reconnaissance de visages, et l'indexation des données vidéo ;
- Le LIFL (Laboratoire d'Informatique Fondamentale de Lille) - TÉLÉCOM LILLE 1, pour la fouille de données, et l'analyse des situations à un niveau sémantique ;
- ARMINES-EMD ( Centre Commun École des Mines de Douai) pour le suivi de trajectoires multiples en temps-réel, et analyse 'bas-niveau' du mouvement dans des vidéos ;
- URECA (UFR de Psychologie, Université de Lille 3) pour l'interprétation des comportements individuels et collectifs ;
- IREENAT (Institut de Recherches sur l'Évolution de l'Environnement Normatif des Activités Transnationales Université de Lille 2) pour l'analyse des problèmes juridiques ;
- Les partenaires industriels, YOUG'S et Thales sont une interface avec les industriels potentiellement intéressés par le projet.